UNIVERSIDADE DE LISBOA

FACULDADE DE PSICOLOGIA

Estudo docimológico dos exames escritos de duas unidades curriculares obrigatórias de um Mestrado Integrado em Psicologia: Psicometria e Psicologia Diferencial

Rita Morais Pequeno Maia

MESTRADO INTEGRADO EM PSICOLOGIA

Secção de Psicologia da Educação e da Orientação

2012/2013

UNIVERSIDADE DE LISBOA

FACULDADE DE PSICOLOGIA

Estudo docimológico dos exames escritos de duas unidades curriculares obrigatórias de um Mestrado Integrado em Psicologia: Psicometria e Psicologia Diferencial

Rita Morais Pequeno Maia

Dissertação orientada pela Professora Doutora Maria João Afonso

MESTRADO INTEGRADO EM PSICOLOGIA

Secção de Psicologia da Educação e da Orientação

2012/2013

AGRADECIMENTOS

À Professora Doutora Maria João Afonso por todos os momentos de aprendizagem, por me ter proporcionado integrar esta investigação, por tudo o que me ensinou, por todos os conhecimentos que comigo partilhou, por todo o apoio, pelo seu precioso *feedback* e pelo seu olhar atento.

À Patrícia pelo caminho que juntas fizemos, por me ter ouvido sempre que precisei, por estar presente incondicionalmente, por acreditar em mim, pelas noites de conversa e de partilhas, por me ter ajudado a crescer, pelo apoio, pela força, pelo carinho e amizade, durante todos estes anos.

Ao António pela paciência nos momentos mais difíceis, pela força e amor incondicional, pois sem ele este percurso não tinha sido o mesmo.

À Inês pela preocupação, por estar presente, por acompanhar de perto este percurso e por ter crescido comigo.

Aos meus irmãos pela preocupação e pelo carinho, por terem aturado as minhas loucuras e por nunca deixarem de estar presentes.

À minha mãe pela ajuda, pelo apoio, pela força, por estar sempre por perto, mesmo estando longe, pelas partilhas que juntas tivemos, por me ajudar a amadurecer, pelo amor incondicional, por todo o esforço que fez para poder alcançar o que hoje tenho.

Ao Tiago Cabaço por ter acompanhado por perto o presente trabalho, pelos recursos e apoio direto e moroso que forneceu. Espero um dia poder contribuir da mesma forma. Um grande obrigada!

Ao André, ao David, ao Nuno, ao Paulo, à Claudia, ao Fernando, ao Sr.º Zé e ao Sr.º Moura por me terem alegrado nos momentos mais "stressantes", mesmo sem saberem.

A todos os colegas e amigos da Faculdade que de alguma forma marcaram este percurso.

RESUMO

A Psicologia Educacional intenta, de forma transversal, a promoção da qualidade do ensino, o que implica que os instrumentos de avaliação das aprendizagens (exames) devam ser objeto de avaliação, por constituírem parte integrante do processo ensino-aprendizagem, particularmente na universidade, onde, por vezes, definem o sucesso ou insucesso e o futuro académico dos estudantes. O presente estudo exploratório, que retoma uma perspetiva clássica de investigação da qualidade dos exames de avaliação de conhecimentos, a docimologia, pretendeu ensaiar uma metodologia de análise passível de aplicação a exames de avaliação das aprendizagens no ensino superior, bem como constituir fundamento para futuras opções relativas aos exames em estudo. Para tal, incidiu na avaliação docimológica dos exames escritos de duas unidades curriculares (u.c.) obrigatórias, do 1ºciclo do Mestrado Integrado em Psicologia (MIP) (Faculdade de Psicologia da Universidade de Lisboa): Psicometria (2010/11) e Psicologia Diferencial (2010/11 e 2011/12). Com base nos resultados de diferentes épocas de exame – num total de 9 exames – este estudo procurou: 1) analisar e avaliar dois formatos distintos de exame, aplicados numa das u.c. (Psicologia Diferencial) e estabelecer comparações entre resultados obtidos em dois anos letivos; 2) comparar os exames de duas u.c. (Psicometria e Psicologia Diferencial), no mesmo ano letivo (2011/12); e 3) analisar longitudinalmente o desempenho de um grupo de estudantes repetentes dessas u.c. A amostra é constituída por 925 estudantes do MIP, cerca de 85% do sexo feminino e 15% do sexo masculino. Os resultados apontam para a vantagem do formato de avaliação utilizado em 2011/12, visto diferenciar eficientemente e avaliar de forma coerente os conhecimentos dos estudantes em ambas as u.c.. A análise dos itens revelou boas qualidades metrológicas destes testes de conhecimentos como também o potencial do método para o estudo de outros exames. Por fim, são assinaladas algumas limitações e implicações para estudos futuros.

Palavras-chave: insucesso escolar; avaliação educacional; avaliação das aprendizagens; docimologia; exames; ensino superior.

ABSTRACT

Educational psychology, through a transversal approach, aims at promoting quality in education, which means that achievement tests (exams) should be evaluated, as they constitute an inherent part of the teaching-learning process, especially at the university, where sometimes they define success or failure and students' academic future. This exploratory study, which recovers the classical perspective of research on the quality of achievement tests, docimology, intended to rehearse a methodology of analysis that can be applied in assessing academic tests in higher education, as well as to establish a basis for future options regarding the academic tests under examination. Therefore, this study focuses on the written exams' docimologic evaluation in two mandatory courses, of the first cycle of a Masters in Psychology (MIP) (Faculty of Psychology, University of Lisbon): Psychometrics (2010/11) and Differential Psychology (2010/11 and 2011/12). Based on the results of different final exams - a total amount of 9 exams - this study has intended to 1) analyze and evaluate two different exam formats, applied to one of the courses (Differential Psychology) and compare the results obtained in two academic years; 2) compare the exams of two courses (Differential Psychology and Psychometrics), in the same academic year (2011/12); and 3) analyze longitudinally the performance of a group of students who failed some of these exams. The sample includes 925 MIP students, about 85% of them female and 15% male. The results demonstrate the advantage of the evaluation format used in 2011/12, as it efficiently discriminates and consistently assesses students' achievements in both courses. The item analysis has revealed good psychometric qualities of these exams and the potential of this methodology in future studies of other academic tests. Finally, some limitations and implications for future studies are pointed out.

Keywords: academic failure; educational assessment; assessment of academic achievement; docimology; exams; higher education.

ÍNDICE

ÍNDICE DE QUADROS DE RESULTADOS

ÍNDICE DE ANEXOS

I. INTRODUÇÃO

As instituições educacionais são fulcrais para o desenvolvimento e bem-estar dos indivíduos, visto desempenharem um papel único para os jovens ao promoverem a compreensão e adaptação ao mundo que os rodeia, para além de uma mais satisfatória vivência em sociedade e uma cidadania mais responsável. Da mesma forma, é na escola que a maioria dos indivíduos adquire conhecimentos, competências, atitudes e experiências que dificilmente conseguiria alcançar de maneira informal noutros contextos. Ao longo do percurso educativo, concretiza-se uma construção conjunta entre professores e alunos, que é única e irreproduzível – o ensino é, assim, um processo de criação (Bernheim & Chauí, 2008).

Têm sido crescentes os esforços da sociedade, no sentido de conferir ao ensino uma boa qualidade, e os investimentos na avaliação têm sido prova disso mesmo, desdobrando-se na avaliação dos professores, das escolas e do sistema de ensino. Assim, tem-se assistido ao aprofundar das investigações e estudos no âmbito da avaliação educacional, que se encontra embutida na própria essência do processo de ensino-aprendizagem, e a vários níveis. Freitas, Sordi, Malavasi e Freitas (2009), na tentativa de sistematizar o campo da avaliação educacional, postulam a existência de três níveis de avaliação, que irão servir de referência ao longo deste trabalho: (1) a avaliação realizada em sala de aula, referente à avaliação das aprendizagens ou avaliação de conhecimentos; (2) a avaliação interna à escola e sob o seu controle, que se refere à avaliação institucional; e (3) a avaliação da responsabilidade de poder público – a avaliação dos sistemas educativos.

Não obstante, apesar dos avanços e desenvolvimentos no que se refere à avaliação, é de assinalar o quanto tem sido descurada a avaliação formal e sistemática dos próprios instrumentos de avaliação, em particular, dos instrumentos utilizados na avaliação das aprendizagens dos estudantes, em sala de aula. Esta perspetiva de investigação, que em certo período do século XX despertou a atenção de investigadores de orientação educacional, diferencial e psicométrica, os quais lhe conferiram o estatuto de domínio de investigação que designaram como docimologia (Piéron, 1974; Miranda,1982; Costa 2007), tem vindo, contudo, a ser negligenciada nos últimos anos. Apesar de os estudos docimológicos terem contribuído de forma inegável para a construção mais fundamentada de testes e exames, a persistência num modelo de avaliação alicerçado no

modelo da medida, levou alguns teóricos a questionar as investigações docimológicas como pertinentes do ponto de vista pedagógico (Leclercq, Nicaise & Demeuse, 2004; Correia, 2002; Despresbiteris, 2009) descurando, assim, a relevância das metodologias e técnicas que a docimologia, como ciência do estudo dos exames, pode promover nas práticas avaliativas conduzidas em sala de aula. Por consequência, assume-se que os métodos de avaliação das aprendizagens, regra geral concebidos informalmente pelos professores, são sempre suficientemente fidedignos e válidos para as tomadas de decisão que suscitam.

Este o problema que inspirou o presente trabalho: a avaliação docimológica no ensino superior. Se a avaliação de aprendizagens é parte integrante e imprescindível do percurso universitário dos estudantes, torna-se incontornável o estudo sistemático dos próprios métodos e instrumentos da avaliação de conhecimentos, por forma a proceder à apreciação da sua qualidade, enquanto instrumentos de medida, e da sua validade para a tomada de decisão relativa ao aproveitamento escolar, com evidentes implicações, não só para o sucesso/insucesso dos estudantes, como para o seu futuro percurso académico e vocacional.

Assim, o presente trabalho incidiu na avaliação docimológica dos exames escritos de duas unidades curriculares obrigatórias, do 1ºciclo (3ºano) do Mestrado Integrado em Psicologia, ministrado na Faculdade de Psicologia da Universidade de Lisboa: Psicometria (ano letivo de 2011/12) e Psicologia Diferencial (anos letivos de 2010/11 e 2011/12). Teve por objetivo ensaiar experimentalmente uma metodologia de análise passível de posterior aplicação a exames escritos semelhantes, de qualquer outra unidade curricular, bem como constituir fundamento para futuras opções relativas à estrutura e tipos de conteúdos dos exames das referidas unidades curriculares.

II. FUNDAMENTAÇÃO TEÓRICA

1. A PSICOLOGIA EDUCACIONAL, OS CONCEITOS DE SUCESSO E INSUCESSO ESCOLAR.

A preocupação com assuntos atualmente enquadrados na Psicologia Educacional remonta ao tempo da Grécia antiga, onde filósofos como Platão e Aristóteles refletiram sobre temas como o papel do professor, a relação entre professor e aluno, os métodos de ensino, a natureza das aprendizagens e o papel das emoções no processo da aprendizagem (Hoy, 2002). Estas mesmas preocupações manifestaram-se, por ocasião da emergência da Psicologia Geral, através do ênfase dado a conceitos como a educação e a formação de professores, debate no qual William James (1842-1910) e Edward Thorndike (1874-1949) tiveram um papel fundamental (Hoy, 2002). Nos anos 60, começaram a surgir divergências entre os próprios psicólogos educacionais, relativamente às diferentes abordagens destes problemas, levando ao aprofundar de teorias sobre o ensino e a aprendizagem. Ainda assim, o campo de estudo da Psicologia Educacional encontrou diversas adversidades que colocaram em causa a construção da sua identidade como ramo da Psicologia distinto e munido de especifidades.

Tal como Wittrock (1992, p.129) expõe, a psicologia educacional é mais do que a convencional definição de "aplicação dos ramos da psicologia à educação", consiste, sim, no estudo científico da psicologia no âmbito da educação. Segundo este autor, a psicologia educacional deverá ocupar-se dos problemas da educação, que incluem o ensino, os processo cognitivos e afetivos dos alunos, o autoconceito, o conhecimento anterior e os conceitos pré-concebidos, o desenvolvimento intelectual e da personalidade, a avaliação educacional, a medição e *o testing* e a orientação vocacional e profissional. Deverá também existir um maior foco (uma verdadeira aposta) no desenvolvimento de métodos quantitativos e qualitativos, no *design* da investigação e técnicas de análise, pois só assim a psicologia educacional poderá contribuir de forma construtiva e rica para a melhoria dos processos de ensino-aprendizagem (Wittrock, 1992).

A psicologia e a educação estão íntima e inevitavelmente articuladas (Miranda, 1982), uma vez que a escola, e tomando em particular atenção, o ensino superior, tem como objetivo primordial a formação de profissionais competentes, criativos e críticos.

Como tal, as intervenções psicológicas no contexto educativo, almejam o desenvolvimento das potencialidades dos sujeitos e a promoção do sucesso escolar (Bisinoto, Marinho & Almeida, 2010).

Assim, a psicologia educacional, entre os seus demais papéis, assume um lugar importante na promoção da qualidade do ensino, sendo o sucesso e insucesso escolar temas importantes, abordados de forma sistemática por psicólogos escolares. O insucesso escolar, em especial, é um conceito aplicado na compreensão do processo de ensino-aprendizagem, remetendo, geralmente, para o fraco rendimento escolar dos alunos.

Etimologicamente, a palavra insucesso deriva da palavra latina *insucessu(m)*, significando *malogro*, isto é, mau êxito, falta de sucesso, fracasso ou desastre. Em Portugal, não existe uma unidade semântica na definição de insucesso escolar, de forma que é relevante referir a análise semântica realizada por Benavente (1990), que através da compilação de vários estudos reuniu para corresponder a esta designação alguns termos, nomeadamente: reprovação, repetência, abandono, desperdício, desadaptação, desinteresse, atraso, desmotivação, alienação e fracasso.

Benavente (1990) refere três teorias explicativas do insucesso escolar: a teoria dos "dons", a teoria do *handicap* sociocultural e a teoria socioinstitucional. Até ao final dos anos 60, a teoria dos "dons" dominou, explica o insucesso escolar mediante as aptidões, ou seja, o sucesso ou insucesso é explicado pelas maiores ou menores capacidades dos alunos – refletindo os seus dotes naturais (Benavente, 1990). A teoria do *handicap* sociocultural, que apareceu no final dos anos 60, explica o insucesso à luz do nível sociocultural do aluno, e o sucesso/insucesso é justificado pela pertença social do aluno a um grupo mais ou menos favorável. Após os anos 70, surge a teoria socioinstitucional, que destaca o papel da instituição para a compreensão do insucesso do aluno (Benavante, 1990), perspetiva que ainda hoje sobressai e é alvo de investigações.

No âmbito do ensino superior, alguns levantamentos de dados têm contribuído para lançar alguma luz sobre os fatores do sucesso/insucesso dos estudantes. Num estudo de Curado e Machado (2005) sobre os percursos escolares dos estudantes da Universidade de Lisboa, onde um dos primordiais objetivos era o estudo dos fatores de sucesso e insucesso escolar, estes autores determinaram que os fatores que influenciam de forma negativa, levando consequentemente a taxas de insucesso mais elevadas, estavam

maioritariamente relacionados com as expetativas dos alunos em relação ao curso selecionado e os apoios encontrados nas instituições. Leandro de Almeida (2004) noutro estudo, conclui que o rendimento académico dos alunos no 1ºano da universidade, se encontra associado especialmente à sua nota de ingresso no ensino superior, bem como às expetativas iniciais e aos comportamentos de envolvimento nas atividades curriculares, vocacionais, institucionais e sociais (cit. por Curado & Machado, 2005).

Não obstante, um dos elementos essenciais que tem um papel central nesta problemática, é a avaliação educacional uma vez que dela depende, em última análise, o sucesso e o insucesso escolar. Luckesi (2002) afirma que a forma como se avalia é crucial para a concretização do projeto educativo, visto que é através da avaliação que os alunos tomam conhecimento do que o professor e a escola valorizam, bem como da qualidade da sua progressão em termos de aprendizagem. Esta deverá ocorrer de forma contínua e sistemática e assumir-se como um procedimento útil na regulação do processo de ensino-aprendizagem dos alunos, fornecendo informações que permitam, aos alunos e aos professores, tomar decisões em tempo útil. No caso de resultados negativos, a avaliação permite propôr medidas no sentido de ultrapassar situações desvantajosas (Dias, 2011).

Esta irá ser uma temática importante a ser abordada, ao longo desta monografia, com o intuito de compreender o papel da avaliação educacional no insucesso escolar. Entramos, então, no domínio da avaliação educacional que abarca uma série de pressupostos e níveis que deverão ser explicitados.

2. A AVALIAÇÃO EDUCACIONAL

2.1. DELIMITAÇÃO DO CONCEITO

Descrever um conceito de tamanha complexidade torna-se uma tarefa claramente difícil. Não só pela extensa gama de variáveis que a avaliação abarca, desde aspetos sociais, económicos, políticos até aos aspetos metodológicos (Despresbiteris,1998) passando pelas opiniões e controvérsias de quantos se vêem envolvidos no processo de ensino-aprendizagem.

Valadares e Graça (1998, p.34) ampliam o conceito afirmando que a "avaliação é uma necessidade vital do ser humano", uma vez que toma várias formas no quotidiano, orientando as decisões, das mais simples às mais complexas. O ser humano utiliza

sistematicamente o julgamento avaliativo para tomar uma decisão de forma válida, na maioria das vezes sem tomar consciência ou compreender o processo que utilizou. Desta forma, a polissemia deste conceito deve-se sobretudo ao seu caráter multidimensional (Valadares & Graça, 1998).

Posto isto, a "avaliação desempenha um papel fulcral em toda a experiência educativa", sendo um fator decisivo na educação (Valadares e Graça, 1998, p.12), uma vez que permite conhecer o aluno, a sua evolução, a consolidação das suas aprendizagens, para além de fomentar experiências educativas posteriores. Podemos, ainda, encontrar algumas definições nos documentos legais em vigor que definem a avaliação como um elemento integrante e regulador da prática educativa, constatando o poder fulcral que assume no processo educativo (Decreto-Lei nº139/2012, de 5 de julho e Despacho normativo nº24-A/2012, de 6 de dezembro).

Nevo, Alkin e Cartstensen (1975) definem avaliação educacional como um processo de recolha de informação, de forma sistemática, tendo em conta a natureza e a qualidade dos objetivos educacionais. Segundo estes autores, a palavra avaliar, encontra as suas raízes na França Antiga, nas palavras *value* e *valoir*, e do latim *valére*, que significa "ter valor" ou "que calcula o valor". Nevo, Alkin e Cartstensen (1975) acrescentam ainda que a avaliação educacional, apesar de ter pontos em comum com outras formas de avaliação, detém caraterísticas particulares, nomeadamente, o facto de as suas raízes se encontrarem na avaliação e medição das aprendizagens dos alunos, o forte envolvimento da sociedade, na prática e no uso da avaliação, e o papel dos professores, que não pode ser dissociado dos resultados das avaliações.

Estima (2011, p.8) refere que, atualmente, estamos perante uma visão holística e sistémica, onde a avaliação ocupa um lugar central na política educativa. Nesta perspetiva, esta constitui "um mecanismo que permite aferir sobre a qualidade das aprendizagens, sendo um instrumento que visa o sucesso educativo". Podemos então admitir que a avaliação é um processo de comunicação social, que orienta o currículo e a prática pedagógica da escola. Tal como referem Albuquerque e Oliveira (2012, p.27) "é uma força criadora do aprender e do ensinar (...), uma declaração de compromisso com a aprendizagem dos alunos", visando assim a compreensão do processo ensino-aprendizagem e a sua concretização.

Pacheco (1995) destaca quatro grandes funções da avaliação educacional: a pedagógica, a de controlo, a crítica e a social. A função pedagógica poderá subdividir-se em quatro dimensões: a dimensão pessoal, ligada à motivação; a dimensão didática, de seleção de métodos e meios adequados à aprendizagem; a dimensão curricular, relacionada com as contextualizações dos currículos e dos programas; e a dimensão educativa, centrada na avaliação do sistema educativo. Assim, podíamos dizer que a função pedagógica funciona como "o barómetro da qualidade do sistema educativo" (Pacheco, 1995, pp. 21). Por outro lado, aponta a função de controlo, como a que é exercida pelo professor de uma forma dissimulada aquando da sua intervenção, e a função crítica relacionada com a melhoria que a avaliação pode promover no sistema educativo. Por último, a função social contempla uma forma de certificação das competências adquiridas pelos alunos (Pacheco, 1995). Hadji (1994) e Estima (2001), por seu lado, distinguem três funções da avaliação das aprendizagens: prever e orientar o processo de ensino-aprendizagem, com base na avaliação diagnóstica; regular e facilitar a aprendizagem, a partir da avaliação formativa; e, por último, certificar e controlar a aprendizagem, através da avaliação sumativa. Assim, compreendemos que a avaliação constitui uma operação indispensável de qualquer sistema escolar, que acompanha o progresso do aluno, ao longo do seu percurso de aprendizagem (Ribeiro, 1991).

Considerando a avaliação educacional uma constante no percurso escolar de qualquer aluno, esta aparece "aliada às aprendizagens realizadas pelos alunos, aos programas das disciplinas, à qualidade do ensino, aos estabelecimentos e ao sistema de ensino" (Afonso, 2011, p.7). Assim, compreendemos que a avaliação educacional, enquanto domínio, apresenta vários níveis, para a presente investigação, iremos dar mais ênfase à avaliação das aprendizagens, ou dos conhecimentos, adquiridos pelos alunos em sala de aula.

No âmbito escolar, são propostas quatro modalidades de avaliação das aprendizagens que importa referir: a avaliação diagnóstica, que averigua a posição inicial do aluno face a novas aprendizagens que lhe vão ser propostas ao longo do processo de ensino-aprendizagem, permitindo antecipar e prevenir dificuldades futuras; a avaliação formativa, que intenta determinar a posição do aluno ao longo de uma unidade de ensino, identificando dificuldades e fornecendo soluções; a avaliação contínua, que poderá ser definida como uma avaliação formativa de caráter permanente;

e a avaliação sumativa que pretende ajuizar o progresso do aluno, no final de uma unidade de aprendizagem, correspondendo, assim, a um balanço final da aprendizagem do aluno (Ribeiro, 1991; Valadares & Graça, 1998; Fernandes, 2011; Zeferino & Passeri, 2007).

Santos e Varela (2007) suplementam uma perspetiva interessante, defendendo que a avaliação das aprendizagens deverá incluir uma dimensão diagnóstica, para que conduza a um melhor ajuste do processo de ensino-aprendizagem. Deste modo, o processo avaliativo deverá percorrer um trajeto que conflua na promoção e consolidação das aprendizagens, pois a necessidade de avaliar será um tema sempre atual e permanente, no contexto escolar. Fernandes (2011) acrescenta que a avaliação deverá estar ao serviço das aprendizagens, não se separando do ensino e dos processos inerentes, com o intuito de auxiliar os alunos a melhorar as suas aprendizagens.

De acordo com Estima (2011), em Portugal, reconhece-se a relevância de articular os objetivos e funções da avaliação formativa e da avaliação sumativa. Deste modo, a avaliação integra e regula as práticas pedagógicas, mas assumindo, paralelamente, uma função de certificação das aprendizagens realizadas. Concluindo, a avaliação das aprendizagens descreve, então, conhecimentos, atitudes ou aptidões que os alunos adquiriram, compreendendo que objetivos do ensino os alunos já alcançaram, num determinado ponto do percurso, e quais as suas dificuldades (Ribeiro, 1991).

A avaliação educacional, e em particualr a avaliação das aprendizagens, é um elemento integrativo e regulador da prática educativa, e do sucesso e qualidade do ensino. Sabemos, porém que diversas mudanças culturais, sociais, históricas, políticas e metodológicas deram origem a diferentes modelos de aprendizagem, e consequentemente a diferentes formas de abordar a avaliação. Importa, então, através de uma breve contextualização histórica, compreender a evolução da avaliação no campo educacional.

2.2. EVOLUÇÃO DO CONCEITO DE AVALIAÇÃO EDUCACIONAL (E DAS APRENDIZAGENS)

Os processos de avaliação constituem, desde há muito, uma preocupação das sociedades humanas. Podemos voltar atrás no tempo e relembrar, por exemplo, as cerimónias de iniciação das tribos primitivas, onde os jovens, para alcançarem um novo

estatuto na sociedade, tinham que superar uma série de desafios, de testes de resistência e de conhecimentos de costumes tribais (Valadares & Graça, 1998). Eram tentativas deveras primárias e pouco fiáveis, mas constituíam uma forma de avaliar e de distinguir os jovens que conseguiam ultrapassar as provas, dos que não eram bem-sucedidos.

Guba e Lincoln (1989) apresentam uma perspetiva organizada que engloba abordagens, significados e conceptualizações, ao longo do século XX, reconhecendo quatro gerações da avaliação educacional.

A geração da medida

A primeira geração, conhecida como a "geração da medida", parte do pressuposto de que a avaliação e a medida são sinónimos, isto é, a avaliação era entendida como uma questão técnica, posto que mediante testes bem construídos, era possível avaliar (quantificar), com rigor e precisão, as aprendizagens escolares dos alunos. Guba e Lincoln (1989) definem dois fatores que influenciaram esta primeira geração da avaliação. O primeiro está relacionado com uma questão de afirmação dos estudos sociais e humanos que se começavam a realizar em Inglaterra, nos Estados Unidos, na Alemanha e em França. A investigação em ciências sociais era aconselhada a seguir o método experimental, no sentido de se afirmar junto da comunidade científica, ganhando credibilidade (Stuffebeam, Madaus & Kellaghan, 2000). Assim, os instrumentos destinados a medir as aprendizagens humanas, que permitiam quantificá-las, compará-las ou ordená-las numa escala (Fernandes, 2004), tornavam possível a quantificação das aprendizagens dos alunos, possibilitando a aplicação do modelo científico, que constituía à época um marco significativo de sucesso, obtendo assim a credibilidade desejada. O outro fator influente, na primeira geração, foi a emergência do movimento da gestão científica no mundo da economia (Guba & Lincoln, 1989). A revolução industrial impôs múltiplas transformações na organização social, levando à necessidade de permanentes avaliações das estruturas existentes (Valadares & Graça, 1998). O que se procurava era tornar mais eficiente, eficaz e produtivo o trabalho dos seres humanos, "colocar a pessoa certa no local certo". Estas avaliações constituíram a base de uma abordagem empírica da avaliação de programas educacionais, estimulando a emergência de contributos importantes para o desenvolvimento de instrumentos de avaliação como Horace Mann (1796-1859), Joseph Rice (1857-1934), e Hermann Ebbinghaus (1850-1909) (Valadares & Graça, 1998). Segundo Guba e Lincoln (1989) a sistematização, a estandardização e a eficiência caracterizam o essencial deste

movimento, onde Frederick Taylor (1856-1915) era o principal teórico. As conceções primordiais do Taylorismo foram rapidamente adotadas pelos sistemas educativos, passando os mesmos a ser percecionados como análogos às organizações empresariais.

Posto isto, nas primeiras décadas do século XX, a avaliação vista como sinónimo de medida foi tão disseminada que se criaram associações e comités encarregados do estudo e elaboração de testes padronizados (Cerny & Ern, 2001), atribuindo um caráter instrumental ao processo avaliativo. Assim, nasce a docimologia, por volta dos anos 20, significando o estudo sistemático dos exames e do sistema de atribuição de notas (Piéron, 1974). Neste percurso, a avaliação foi orientada pelos estudos docimológicos "restrita ao estudo dos exames e fundada no modelo da medida ou modelo psicométrico" (Cerny & Ern, 2001, p.2). Esta temática será abordada de forma mais extensa adiante, neste trabalho.

De um modo sistemático, a *geração da medida* baseava-se numa perspetiva onde prevaleciam as funções sumativa, classificativa e seletiva da avaliação, sendo o único objeto da avaliação os conhecimentos dos alunos (pouco ativos no processo), levando a uma avaliação descontextualizada, onde se privilegia a quantificação das aprendizagens (Fernandes, 2004).

A geração da descrição

A segunda geração tentou superar algumas das limitações entretanto detetadas. Desta forma, os avaliadores, perante objetivos educacionais previamente definidos, tinham como principal objetivo descrever padrões de pontos fortes e fracos (Guba e Lincoln, 1989). É assim definida como "geração da descrição", uma vez que não se limitavam a medir, mas procuravam, sim, descrever se os alunos atingiam os objetivos definidos *a priori*. Desta forma, a medida deixou de ser um mero sinónimo de avaliação, mas tornou-se num dos veículos ao seu serviço (Guba e Lincoln, 1989). Ralph Tyler é referido como tendo uma influência significativa nesta geração (e até 1965) nos sistemas educativos, sendo pioneiro no desenvolvimento da perspetiva de formulação de metas para melhor se definir o objeto de avaliação, trazendo uma nova visão do currículo e da avaliação (Valadares & Graça, 1998; Guba & Lincoln, 1989). É com este autor que nasce a expressão "avaliação educacional", com o intuito de designar o processo de avaliar em que medida os objetivos eram ou não alcançados no sistema educativo.

A grande diferença em relação à geração anterior está em que ao se formularem objetivos comportamentais e se verificar se estão a ser atingidos pelos alunos, a avaliação é caracterizada primordialmente pelo desenvolvimento de uma "função reguladora" e da "preocupação em conceptualizar o currículo de forma abrangente" (Fernandes, 2004, pp.11).

A geração da formulação de juízos ou julgamentos

A terceira geração, designada por Guba e Lincoln (1989) como a "geração da formulação de juízos ou julgamentos", nasce da necessidade de superar falhas da geração anterior. De acordo com o postulado por Guba e Lincoln (1989), os avaliadores passariam a desempenhar o papel de juízes, fazendo esforços para que as avaliações permitissem formular juízos de valor acerca do objeto de avaliação. Também é de referir que durante este período (anos 50-60) se assistiu a uma significativa disseminação de programas educacionais e respetiva avaliação, surgindo também as primeiras taxonomias de objetivos educacionais como as de Bloom (1956) e Guilford (1959) (Valadares & Graça, 1998). Esta geração ficaria igualmente marcada pelo lançamento do Sputnik, em 1957, pela União Soviética, que levou a que o Ocidente desenvolvesse profundas reformas educativas com o intuito de promover essencialmente o ensino da matemática e das ciências, com receio de que ficasse para trás no desenvolvimento científico e tecnológico (Guba & Lincoln, 1989).

Foi, assim, uma época de grande expansão e desenvolvimento da avaliação, a qual alguns autores denominam de "Idade de Desenvolvimento" (Stuffebeam, Madaus & Kellaghan, 2000; Valadares & Graça, 1998). Outro marco importante foi a distinção entre o conceito de avaliação sumativa e formativa, por Michael Scriven, em 1967 (Fernandes, 2004).

De forma sucinta, esta geração desenvolve conceitos importantes e determinantes para a evolução da avaliação educacional, implantando a ideia de que o processo de avaliação deve facilitar a tomada de decisões, deve envolver todos os agentes do processo educativo (pais, professores e alunos), deve tomar em consideração os contextos de ensino e aprendizagem e deve definir critérios de apreciação dos testes/exames (Fernandes, 2011).

A geração da negociação e construção

Guba e Lincoln (1989) propõem, por fim, a geração de rutura epistemológica com as anteriores, a "geração de negociação e construção". A quarta geração caracteriza-se por não estabelecer, *a priori*, parâmetros ou enquadramentos, pois estes serão determinados ao longo de um processo negociado e interativo com todos os envolvidos na avaliação (Guba & Lincoln, 1989). Por um lado, trata-se de uma avaliação construtivista, que segundo Fernandes (2011, p.13), está baseada num conjunto de princípios, entre os quais se destacam os seguintes: a avaliação como conceito relativo, dependente de quem o faz e de quem nela participa; os professores deverão partilhar, com os alunos, o poder de avaliar; o *feedback* é um elemento indispensável na avaliação; e a avaliação deverá ajudar os alunos a desenvolver as suas aprendizagens. Assim, a função formativa da avaliação é colocada em destaque, e o cerne desta nova geração passa por o aluno ter um papel mais ativo no processo de ensino-aprendizagem, sendo a avaliação das aprendizagens o instrumento condutor do desenvolvimento da aprendizagem, das capacidades e das competências. Segundo Rosales (1992), nesta ótica, a avaliação tem como principal finalidade a melhoria qualitativa da educação para que os alunos usufruam de igualdade de oportunidades e desenvolvam atitudes, competências e saberes essenciais à sua formação.

Como percebemos, o conceito de avaliação educacional sofreu alterações ao longo do seu percurso na história da psicologia educacional, passando por mudanças epistemológicas e metodológicas, que confluíram no que conhecemos hoje. Partimos de um conceito psicométrico, de avaliação como mensuração das aprendizagens, e evoluímos em direção a um conceito constutivista, holístico e sistémico, que pressupõem a consolidação e melhoria do processo de ensino-aprendizagem, não tendo em conta apenas a apreciação quantitativa das aprendizagens.

2.3. INVESTIGAÇÕES REALIZADAS EM PORTUGAL

Segundo a meta-análise de todas as monografias desenvolvidas no âmbito da avaliação educacional, realizada por Martins (2008), os estudos realizados nos últimos 30 anos dividem-se em diferentes dimensões de análise: estudos que se ocupam da reflexão acerca das mudanças conceptuais na avaliação; estudos que atribuem à avaliação um papel importante na criação de competências de autorregulação e autoavaliação; estudos de caracterização de ambientes de avaliação em sala de aula;

estudos de análise das relações entre a avaliação formativa e sumativa; e, por último, estudos de reflexão sobre as ligações entre a avaliação interna e externa. Martins (2008) verificou que a avaliação é investigada indiretamente, mediante as conceções dos intervenientes (essencialmente os professores), que o nível de ensino mais estudado é o secundário e as metodologias mais utilizadas são a entrevista e os questionários. Não obstante, nas duas últimas décadas tem sido produzido um número elevado de trabalhos de reflexão e de investigação centrados na avaliação das aprendizagens, principalmente em três domínios de análise: políticas educativas, produção de materiais e investigação (Martins, 2008; Fernandes, 2006).

Apesar da atenção dada à avaliação como tópico de reflexão e investigação, há que reconhecer que têm sido escassos os estudos que se dedicam à análise dos procedimentos de construção dos exames, nos mais variados graus de ensino, bem como à apreciação da qualidade metrológica (validade e precisão) das classificações académicas, o que consiste numa lacuna na investigação da avaliação das aprendizagens atual. Esta lacuna é tanto mais grave quanto é do domínio comum o reconhecimento de que os instrumentos de avaliação psicológica, utilizados em contexto educativo, entre outros (testes de inteligência, aptidões, personalidade, interesses, etc.) devem obedecer a rigorosos critérios e procedimentos de construção e submeter-se a exigentes estudos da sua qualidade técnica, muitas vezes obrigando a sucessivos aperfeiçoamentos, antes da sua divulgação para o uso na prática, na tomada de decisão acerca dos indivíduos.

Ainda que se possa afirmar que os exames ou outros instrumentos de avaliação das aprendizagens não são, em rigor, testes de avaliação psicológica, é inegável que com eles partilham algumas características. Enquanto instrumentos de avaliação de conhecimentos, dos quais decorrem decisões (atribuição de classificações, aprovações/reprovações, creditação de formação adquirida, acesso a outros níveis de ensino, escolhas vocacionais, colocação profissional, etc.) com evidentes implicações para o futuro dos sujeitos avaliados, parece razoável afirmar-se que deveriam pautar-se por critérios de construção e avaliação da sua qualidade idênticos aos de outros métodos de avaliação do funcionamento psicológico individual.

Este ênfase no aperfeiçoamento da qualidade dos instrumentos da avaliação sumativa não pressupõe o abandono das orientações atuais da avaliação das aprendizagens, antes visa proporcionar instrumentos mais sólidos e fidedignos ao serviço dessas orientações. Por outras palavras, a adoção de uma ótica da avaliação

educacional de natureza construtivista, holística e sistémica não deveria dispensar a preocupação com o rigor da construção e estudo dos próprios instrumentos de avaliação de conhecimentos, cuja qualidade é indispensável, seja qual for a perspetiva em que seja tomada, enquanto fonte de informação sobre o processo de aprendizagem de cada estudante.

2.4. Instrumentos de Avaliação das Aprendizagens/de Conhecimentos

A principal função dos instrumentos de avaliação de conhecimentos, em sala de aula, é medir o desempenho dos alunos e, apreender de forma concisa os seus conhecimentos, a eficácia dos seus esforços, para além de motivar e direcionar a sua aprendizagem (Ebel & Frisbie, 1986). A construção destes instrumentos é uma das maiores responsabilidades dos professores (Sax, 1980), que podem optar por diversos formatos: testes/exames sumativos, relatórios, *portfolios* e/ou trabalhos expositivos ou dinâmicos. Dado o muito disseminado recurso a testes sumativos, designadamente, exames escritos, sobretudo em contextos com elevado número de estudantes, como no ensino superior, estes serão o alvo de estudo deste trabalho, uma vez que apresentam, para mais, uma maior complexidade de construção, em particular no que diz respeito ao desenvolvimento dos itens.

Os testes de avaliação das aprendizagens (exames escritos) pretendem prioritariamente discriminar diferentes graus de conhecimentos por parte dos alunos (Cortesão, 2005), permitindo averiguar o domínio cognitivo de um conjunto de temáticas e proporcionando uma verificação ampla do conhecimento adquirido (Zeferino & Passeri, 2007). Ribeiro (1991, p. 92) refere que um teste sumativo intenta realizar um balanço sobre as aprendizagens adquiridas, incidindo numa área vasta dos conteúdos; apresenta, assim, uma "estrutura de malha larga (…) sobre uma extensão vasta da matéria". Este tipo de teste, que incide fundamentalmente sobre aspetos cognitivos da aprendizagem, segundo Ribeiro (1991, p. 93) enquadra-se em duas categorias: testes referidos a normas e testes referidos a critérios. Os primeiros foram, durante largos anos, a forma mais popular de avaliar conhecimentos e aprendizagens, levando à seriação dos alunos submetidos ao processo. Na primeira metade do século XX, acreditava-se que os resultados da avaliação em amplos grupos formavam uma distribuição Normal, a denominada curva de *Gauss*, permitindo assim verificar o

desempenho de um aluno comparando o seu rendimento com a distribuição de resultados do grupo de pertença (Ribeiro, 1991; Vianna, 1998; Sax, 1980). Desta forma, os testes corriam o risco de ter um fraco poder discriminativo, uma vez que, para poder distinguir os melhores alunos, os professores incluíam perguntas extremamente difíceis e complexas, a que possivelmente quase ninguém responderia (Sax, 1980). Por outro lado, os testes referidos a critérios permitem interpretar o desempenho do aluno relativamente a um conjunto de competências e objetivos – definido *a priori* como critério mínimo de elegibilidade ou aprovação (Ribeiro, 1991). Isto é, neste tipo de testes, procura-se determinar o domínio do aluno sobre um conjunto de pré-requisitos ou competências, para além de que informam até que ponto os objetivos de uma unidade de ensino foram realmente alcançados (Vianna, 1998). Para tal funcionar de forma adequada, é necessário que os objetivos selecionados constituam uma amostra representativa do universo de objetivos possíveis, ou seja, do currículo da unidade de ensino. Em Portugal, os testes referidos a critérios são os mais sistematizados e utilizados no sistema de ensino (Vianna, 1998; Ribeiro, 1991).

De um modo mais específico, os testes, quer referidos a normas, quer referidos a critérios, podem ser construídos mediante dois grandes tipos de itens, perguntas de resposta curta e perguntas de resposta longa.

2.4.1. Itens de Resposta Curta ou Breve ou do Tipo Objetivo

Neste grupo, podemos distinguir duas categorias: (1) o aluno dá ou completa a resposta, e (2) o aluno seleciona a resposta de entre alternativas que lhe são propostas. Esta distinção é deveras importante, uma vez que se situam em níveis de complexidade diferente, no primeiro tipo o universo de respostas possíveis é ilimitado, e no segundo é limitado (Ribeiro, 1991).

1) Os itens de resposta breve, onde o aluno dá ou completa a resposta, têm por base uma questão à qual o aluno deverá responder de forma curta ou sintética e sem qualquer ambiguidade; utilizam-se principalmente para avaliar definições de termos e conceitos, de factos específicos, de princípios, de conhecimentos de métodos ou procedimentos e a capacidade para interpretar dados simples (Valadares & Graça, 1998; Ribeiro, 1991, Popham, 1975; Sax, 1980; Ebel & Frisbie, 1986). Assim, são de fácil elaboração, permitindo avaliar vários objetivos no mesmo teste, sendo que não permitem ao aluno

que adivinhe a resposta. Por outro lado, apresenta também algumas desvantagens, como o facto de não se adequar a avaliar aprendizagens complexas e fomentarem muitas vezes a aprendizagem mecânica – memorização (Valadares & Graça, 1998; Sax, 1980).

2) As perguntas de escolha múltipla (onde o aluno seleciona uma alternativa de resposta entre as que lhe são propostas) são tipicamente compostas por um tronco ou uma base, com uma questão ou uma afirmação incompleta, à qual se seguem várias opções de resposta, as alternativas de resposta. Desta forma, estas compreendem uma alternativa correta, ou mais correta, e as alternativas falsas, ou menos corretas, que são, em geral, designados de distratores. As vantagens e desvantagens deste tipo de item são concordantes entre muitos autores, nomeadamente, Valadares e Graça (1998), Fernandes (2004), Ribeiro (1991), Ramraje e Sable (2011), Lee, Liu e Linn (2011), De Landsheere (1976), Simkin e Kuechler (2005), Popham (1975). De um modo geral, estes itens permitem abranger uma parte substancial do domínio a avaliar em relativamente pouco tempo, e quando construídos adequadamente permitem avaliar aprendizagens complexas, sendo bastante objetivos, uma vez que não permitem flutuações de resposta. Para além disso, são facilmente compreendidos por alunos de todas as idades, a avaliação e classificação das respostas é extremamente simples e rápida e a probabilidade de o aluno adivinhar a resposta pode ser reduzida aumentando o número de alternativas. Não obstante, apresentam desvantagens: a construção dos itens consome muito tempo, uma vez que a identificação de alternativas falsas (distratores) plausíveis é muitas vezes difícil. E apesar de ser passível de avaliarem aprendizagens complexas, não abrangem a organização de ideias, a integração e relação entre conhecimentos e permitem que os alunos respondam ao acaso.

2.4.2. ITENS DE RESPOSTA LONGA

Quanto aos itens de resposta longa, livre ou orientada (balizada por objetivos específicos), permitem ao indivíduo, a partir de uma questão ou tema, apresentar as suas ideias sobre o assunto e estruturar a resposta. Desta forma, avaliam a expressão escrita, as aprendizagens complexas como a organização e síntese de ideias, a criação de textos originais e criativos, a análise crítica de documentos, a capacidade de resolução de problemas e a integração de múltiplos conhecimentos (Ribeiro, 1991). Como vantagens também podemos referir que a elaboração é rápida e relativamente fácil e é possível analisar os processos e estratégias utilizados pelos alunos na resolução dos problemas,

algo que não é possível a partir de perguntas de escolha múltipla (Fernandes, 2004; Ribeiro, 1991). A maior desvantagem consiste na necessidade de recorrer a critérios de classificação, o que reduz a objetividade da avaliação, verificando-se, por vezes, oscilações entre classificações atribuídas, quer por avaliadores diferentes, quer pelo mesmo avaliador em diferentes momentos (Ribeiro, 1991; Sax, 1980), sendo que a fiabilidade entre juízes tende a ser mais baixa. Para além disso, consomem muito tempo de análise e de classificação, uma vez que é necessário, da parte de quem avalia, o exercício complexo de comparação de cada resposta (e a diversidade de respostas é potencialmente ilimitada) com os critérios de correção bastante objetivos definidos *a priori*. Para mais, tendem a favorecer os alunos com maior facilidade de expressão escrita, mesmo quando não seja essa a aptidão a avaliar, o que reduz a validade da avaliação como medida de conhecimento.

Por serem diversas as vantagens e desvantagens que resultam da utilização de cada tipo de formato, será ideal procurar diversificar o tipo de questões ou de itens dos instrumentos de avaliação, equilibrando e compensando, desse modo, as exigências e as potencialidades dos diversos formatos, quanto à validade e à fiabilidade dos testes (Fernandes, 2011; Ebel & Frisbie, 1986)

3. A AVALIAÇÃO DA AVALIAÇÃO DAS APRENDIZAGENS – A DOCIMOLOGIA

Após refletirmos e analisarmos a interface da psicologia educacional com o estudo do insucesso e do sucesso escolar, e com a avaliação educacional, impõe-se a necessidade de introduzir a avaliação dos próprios instrumentos usados na avaliação das aprendizagens. Tal como foi referido anteriormente, atualmente são escassos os estudos envolvendo a análise dos procedimentos de construção dos testes, bem como de análise da qualidade dos itens, da precisão das classificações ou da validade dos resultados para a tomada de decisão, em testes de avaliação de conhecimentos, realizados em sala de aula, como os teste ou exames escritos. Assim sendo, há que reconhecer a relevância de investigar mais aprofundadamente neste domínio – a *avaliação da avaliação das aprendizagens* – começando por um levantamento do "estado da arte".

Entramos assim, no campo da docimologia, da ciência dos exames. O estudo científico dos problemas psicopedagógicos da avaliação de conhecimentos escolares em situação de exame foi iniciado, de forma sistemática, nos princípios dos anos 20, por

Henri Piéron (Miranda, 1982). Este eminente psicólogo denominou esta área de "docimologia", significando o estudo sistemático dos exames, que emerge da preocupação de que os exames se transformem na única finalidade do ensino, quando deveriam constituir um meio de verificar a sua eficácia (Piéron, 1974). De Landsheere (1976) distingue três conceitos fundamentais nesta área: a "docimologia" como a ciência que estuda sistematicamente os exames, particularmente o sistema de atribuição de notas e os comportamentos dos examinadores; a "docimástica" definida como a técnica dos exames; e a "doxologia", ou estudo sistemático do papel que a avaliação desempenha na educação escolar.

Piéron (1974) iniciou o primeiro estudo docimológico, em 1922, analisando o exame de certificação do ensino primário, conjuntamente com Henri Laugier e Mme. Piéron. Para tal, aplicaram a 117 estudantes, no fim do 4ºano, um grupo de seis testes relativos a capacidades diversas, nomeadamente, troca de letras, formação de palavras, frases absurdas, analogias, memória imediata de palavras e perceção visual de trocas (Piéron, 1974). As classificações finais, relativas ao 4ºano, e as classificações obtidas no exame de certificação foram calculadas, e divididas em três grupos: (1) relativo a aquisições mnemónicas, relacionadas com história e geografia, recitação e ortografia; (2) relativo a capacidades intelectuais, relacionado com redação, leitura expressiva e aritmética; e (3) relativo a qualidades diferentes ou "capacidades extrainteletuais", ligado ao desenho, à caligrafia, ao canto, à ginástica, à apresentação do caderno (Piéron, 1974). Piéron e os seus colaboradores esperavam encontrar uma correlação muito baixa, ou quase nula, entre o último grupo e os outros dois, porém todas as correlações foram de valor médio e bastante semelhantes. Para além disso, foi quase nula a correlação de cada um dos seis testes com as classificações escolares (ao longo do ano letivo e exame final), levando Piéron (1974, p.14) a concluir que os exames de certificação do ensino fundamental constituem um "dado bastante pobre e muito insuficiente (…) não nos podemos limitar a um exame de tipo tradicional, nem mesmo atribuir a essa prova um valor eliminatório decisivo".

A partir deste estudo docimológico, outros trabalhos surgiram, tendo um foco primordial na fiabilidade e estabilidade das notas. Segundo Piéron (1974), o primeiro estudo posto em evidência foi de Laugier e Weinberg: estes investigadores analisaram 166 composições de história e geografia corrigidas por dois examinadores distintos. Constataram que eram evidentes certas divergências, como por exemplo, um candidato

classificado com a segunda melhor nota, por um professor, estava classificado em antepenúltimo lugar, por outro. As diferenças entre os examinadores na correção das provas eram grandes, revelando a inconsistência nos critérios e pondo a nu um fator subjetivo na apreciação das mesmas. Muitos outros estudos foram levados a cabo com o mesmo propósito, Piéron (1974) e Leclercq, Nicaise e Demeuse (2004) citam alguns autores que colocam então em evidência a urgência de se explorar e estudar os problemas de validade e fiabilidade dos exames, como Sadler (1932), Jadoulle (1932), Eels (1930), Laugier e Weinberg (1927), Laugier, Piéron, Piéron, Toulouse e Weinberg (1934).

Posto isto, a docimologia surge numa época de crítica extrema e pouca confiança nos métodos tradicionais utilizados para fins de seleção em exames e concursos (Costa, 2007), emergindo, assim, duas grandes linhas de investigação: a docimologia clássica, também denominada por "negativa", e a docimologia experimental, também intitulada de "positiva" (Miranda, 1982; Costa, 2007). A docimologia clássica enfatizava o aperfeiçoamento das técnicas de avaliação e da análise de resultados, enquanto a docimologia experimental se centrava na análise da avaliação como um comportamento, tentando determinar os mecanismos intervenientes na decisão avaliativa (Costa, 2007). Segundo, Miranda (1982, p.47) a docimologia clássica enquadra-se num "plano de verificação e de análise das divergências da avaliação, em situação natural ou provocada". Essas divergências ocorrem em diversos casos e foram muitas as investigações que as colocaram em evidência, nomeadamente, o acordo interjuízes baixo, relativo às classificações do mesmo aluno na mesma disciplina e à dispersão das classificações atribuídas pelos mesmos juízes e por juízes diferentes (Piéron, 1974; Noizet e Caverni, 1985). Por outro lado, de acordo com Miranda (1982, p. 47) a docimologia positiva, comporta duas linhas primordiais: (1) a que diz respeito ao aperfeiçoamento da avaliação, isto é, "o desenvolvimento de técnicas de construção dos instrumentos de avaliação", como a análise das qualidades psicométricas e o estudo de adequação; (2) e a "docimologia experimental, que se centra na avaliação como um comportamento" (p.48), ou seja, uma forma global de responder a uma situação na qual nos encontramos. Procura, portanto, determinar experimentalmente os mecanismos cognitivos e percetivos intervenientes na decisão avaliativa, bem como os fatores de distorção presentes no processo avaliativo (Noizet & Caverni, 1985).

Assim sendo, os estudos docimológicos puseram em evidência as divergências da avaliação mediante comparações sistemáticas, em situações naturais e provocadas, resultando destes trabalhos estratégias de redução das divergências. Na sua excecional monografia "A docimologia em perspetiva", Miranda (1982) realiza uma descrição detalhada dessas estratégias e avanços realizados, com o intuito de aumentar a validade e precisão das avaliações, e tornando as classificações dos exames comparáveis, exprimindo as diferenças individuais entre os alunos e não entre os examinadores. Paralelamente, Miranda (1982) equaciona a investigação docimológica como um contributo importante, traçando o seu caminho de uma perspetiva crítica aos processos de avaliação, que põe em evidência a instabilidade das classificações e as divergências entre os examinadores, para uma perspetiva de estudo científico dos exames e das técnicas da avaliação das aprendizagens. Os contributos que a docimologia trouxe à avaliação das aprendizagens foram muitos, por um lado pela pertinência das questões avaliativas e, por outro, pela reformulação da problemática da avaliação educacional na perspetiva de realização de objetivos pedagógicos.

Não obstante, a docimologia dita clássica, referente ao estudo sistemático dos exames, segundo alguns autores, desvalorizava a avaliação, uma vez que pressupunha o exame como processo exclusivo da avaliação (Correia, 2002; Despresbiteris, 1998; Leclercq, Nicaise & Demeuse, 2004). Deste modo, apesar dos avanços teóricos e metodológicos alcançados, a avaliação de conhecimentos objetivava primordialmente a aplicação e análise de testes, atribuindo ao processo avaliativo um caráter meramente instrumental (Boavida, 1985, cit. por Fernandes, 2006). Denotou-se, então, uma maior aposta nos estudos enquadrados na docimologia positiva, de onde emerge a análise experimental do comportamento da avaliação. Miranda (1982, p.56) expõe os vários estudos realizados neste âmbito, referindo que, nesta perspetiva, o estudo das divergências da avaliação contém "aspetos de natureza cognitiva e percetiva, e determinados pela interação entre as variáveis do estímulo e as variáveis da personalidade".

Posto isto, o estudo da avaliação passa de uma perspetiva quantitativa, de análise metodológica e psicométrica, para um estudo mais qualitativo, com ênfase na análise da situação de avaliação e nas interações resultantes. Como tal, podemos equacionar os problemas da avaliação de conhecimentos, resultantes dos estudos docimológicos, em duas fases: a docimologia clássica (ou negativa) tornou evidente a instabilidade das

classificações dos exames e as diferenças inter e intraindividuais dos examinadores na apreciação dos exames; e a docimologia experimental (ou positiva) expôs a diversidade dos critérios de apreciação e de classificação, bem como os fenómenos de interação da situação avaliativa com o indivíduo, como as expectativas *a priori* subjacentes à apreciação (efeito de *halo*), ou os fenômenos de ancoragem em que, por exemplo, a ordem de apresentação das provas e a nota imediatamente anterior, influenciam a apreciação das provas posteriores provocando distorções, como sobrevalorização ou efeitos de contraste, entre outras (Miranda, 1982; Noizet & Caverni, 1985; Chabot, 2004).

Verificamos então que os estudos no âmbito da docimologia clássica, apesar de pertinentes, foram abandonados e descurados nas investigações no domínio da avaliação das aprendizagens, resultando num maior ênfase dado aos estudos qualitativos. E com os avanços realizados na avaliação das aprendizagens, atualmente os instrumentos utilizados têm por base um processo formativo, que descura, em grande parte, o valor instrumental dado à avaliação. Contudo, existe um maior número de publicações que apresentam normas e critérios, com o intuito de auxiliar na construção de testes e/ou exames escritos, de modo obter uma maior qualidade, pelo que se torna pertinente retornar aos estudos docimológicos, para analisar de forma sistemática a qualidade destes instrumentos de avaliação de conhecimentos, utilizados em grande escala no ensino superior, com uma função, muitas vezes, determinante do sucesso/insucesso académico.

4. A AVALIAÇÃO DA APRENDIZAGEM NO ENSINO SUPERIOR.

Após a conceptualização do campo teórico e metodológico que enquadra o presente estudo, torna-se essencial estreitar o domínio para o nível de escolaridade pertinente para a investigação – o ensino superior, percebendo a função da avaliação das aprendizagens neste grau de ensino e os seus propósitos.

Segundo Garcia (2009, p. 205), diversos são os estudos sobre a avaliação da aprendizagem na educação superior, que sugerem a "existência de uma relação estreita entre as práticas de avaliação desenhadas pelos professores e os níveis de desenvolvimento dos estudantes". As experiências avaliativas podem influenciar o modo como os estudantes planeiam e utilizam o seu tempo de estudo, como atribuem prioridades e significado às diferentes tarefas académicas, e simultaneamente

influenciar a própria aprendizagem adquirida pelos alunos ao longo da unidade de ensino, sendo assim, a avaliação um instrumento central no processo ensino-aprendizagem (Garcia, 2009; Struyven, Dochy & Janssens, 2005; Rehem & Melo, 2008).

O foco primordial da avaliação, no ensino superior, deverá ser a formação para a vida em sociedade, portanto desenvolver, afirmar, consolidar conhecimentos, competências e atitudes nos estudantes, tendo em vista a futura produção e transmissão de conhecimento. A avaliação é um processo de formação, de apropriação dos sentidos das experiências, das situações e dos projetos de vida (Sobrinho, 2010). A avaliação pretende mensurar o conhecimento, verificar se o aluno se encontra apto em determinada unidade de ensino. Deste modo, a avaliação é um momento de aferir a aprendizagem, através de testes, provas ou/e trabalhos, que serão consideradas como produto final de um semestre (Santos, 2012). Assim, o professor na avaliação de conhecimentos no ensino superior, tem um papel de mediação da aprendizagem do aluno, sendo potenciador de reflexões e mudanças na aprendizagem.

A avaliação no ensino superior, isto é, a medição do nível de desempenho escolar e a atribuição das classificações, faz-se por meio de instrumentos cuja qualidade, em geral, não é avaliada (Bittencourt, Creutzberg, Rodrigues, Casartelli & Freitas, 2011). Para além disso, segundo Santos (2012), o espaço dado à discussão e análise sobre a avaliação no ensino superior é pequeno em comparação com outros níveis de ensino. Assim, torna-se imprescindível a realização de estudos sistemáticos que remetam para a análise e reflexão das práticas avaliativas, neste nível de ensino, primordialmente no que diz respeito à análise da qualidade técnica dos instrumentos (análise dos itens, fiabilidade ou precisão e validade), posto que tais instrumentos são decisivos na determinação do sucesso/insucesso académico, bem como na certificação de conhecimentos e competências adquiridos.

4.1. A Avaliação nas duas Unidades Curriculares sob Análise.

Após a revisão de literatura, e a análise dos estudos e investigações no âmbito desta temática, a avaliação de conhecimentos emerge como um elemento integrativo e central no processo de ensino-aprendizagem. Posto isto, impõem-se como imprescindível a procura de soluções, para que a sua função no processo pedagógico e educativo seja

promotora de uma aprendizagem significativa, refletindo um percurso mútuo de aprendizagem, onde o formando poderá apreender os seus limites e potencialidades e o professor aprimorar os seus métodos de ensino e de avaliação.

É assim que, como parte integrante das preocupações pedagógicas na Faculdade de Psicologia, da Universidade de Lisboa, instituição formadora de psicólogos educacionais que, na sua formação, incluem inevitavelmente um processo de ensino-aprendizagem, com vista à promoção da educação e do ensino, se afigura fulcral a reflexão e análise desta temática. Nesta ótica, propõe-se neste estudo a análise dos instrumentos de avaliação escrita de conhecimentos (exames), utilizados em duas unidades curriculares obrigatórias do currículo da formação no 1ºciclo do Mestrado Integrado Psicologia (3ºano): Psicometria (1ºsemestre) e Psicologia Diferencial (2ºsemestre).

A Psicometria e a Psicologia Diferencial (Psi. Diferencial), enquanto unidades de ensino, pretendem potenciar o desenvolvimento de competências e atitudes epistemológicas, em investigação fundamental e aplicada, relativamente a modelos e a metodologias de observação psicológica. De forma mais específica, a Psicometria tem por intuito a aquisição de conhecimentos sobre a teoria psicométrica, os fundamentos teóricos e empíricos da medição psicológica, a construção e estudo metrológico de medidas e a aquisição de competências de aplicação de metodologia psicométrica de avaliação. Paralelamente, a Psi. Diferencial intenta a aquisição de conhecimentos científicos, de competências técnicas e deontológicas, no domínio da abordagem diferencial, bem como o desenvolvimento de ferramentas de concetualização das diferenças psicológicas individuais, em dimensões, variedades e fatores de diferenciação.

Estas duas unidades curriculares percecionam-se como importantes, na formação de psicólogos, uma vez que têm em vista a aquisição de conhecimentos, de competências e de ferramentas para a aplicação e compreensão de metodologias de investigação, tanto num domínio psicométrico como diferencial. Particularmente, a aquisição de conhecimentos psicométricos possibilita o domínio e uso adequado dos métodos utilizados em avaliação psicológica (testes psicológicos) transmitindo conhecimentos sobre os fundamentos teóricos e os métodos e técnicas de construção, avaliação e utilização desse tipo de instrumentos. Os conhecimentos adquiridos em Psi. Diferencial pretendem, por seu lado, garantir uma preparação de base epistemológica teórica e

metodológica, no domínio do estudo das diferenças intra e interindividuais e inter grupais. Ambas as unidades curriculares permitem uma análise reflexiva sobre a evolução dos paradigmas de investigação, das teorias e metodologias, bem como a tomada de consciência da sua aplicabilidade à avaliação e intervenção em Psicologia. Para além disso, são debatidos os princípios deontológicos e éticos importantes para a intervenção do psicólogo no âmbito da investigação diferencial e, muito em particular da avaliação em psicologia.

A constatação de que nestas unidades curriculares, existia uma taxa de insucesso escolar significativa, levou as docentes a adotarem novos formatos de avaliação de conhecimentos, com o intuito de melhor apreender o nível de conhecimentos adquiridos pelos estudantes, e diversificar as técnicas de avaliação, tornando-as adequadas a uma maior variedade de estudantes e promover a qualidade do ensino. De facto, por muitos anos, a avaliação escrita nestas unidades curriculares era efetuada através da resposta a temas para desenvolvimento (item de resposta longa), um para a avaliação de conhecimentos de matéria teórica e outro de matéria prática. Esta prática, para além de ser exigente do ponto de vista da classificação das respostas abertas, tende, eventualmente, a favorecer os estudantes com maior capacidade de expressão linguística escrita. Por outro lado, o reconhecimento de que é desejável a diversificação dos formatos de avaliação de conhecimentos, com recurso a itens de diferentes tipos, levou à conceção de um novo formato de exame, impondo-se em consequência a necessidade de ensaio empírico do novo formato.

Caracterizando brevemente os instrumentos de avaliação de conhecimentos das unidades curriculares acima descritas, tratam-se de testes referidos a critérios (Ribeiro, 1991), constituídos por dois tipos de perguntas: itens de escolha múltipla (três alternativas de resposta) com solicitação de justificação (em que o aluno deve justificar e esclarecer o fundamento da escolha da alternativa, numa formulação breve, de cinco linhas); e tema (s) de desenvolvimento, em que perante cada tema, o estudante deverá elaborar uma resposta articulada e sintética, com a extensão aproximada de duas páginas. Uma vez que o presente trabalho incide nos exames de dois anos letivos, os primeiros anos em que se ensaiou este novo formato de exame com dois tipos de questões, serão analisados exames com duas estruturas diferentes: subdivisão “Parte Teórica/Parte Prática” (cinco itens de escolha múltipla e um de desenvolvimento em

cada parte) ou subdivisão "Escolha Múltipla/Desenvolvimento" (dez itens de escolha múltipla e um de desenvolvimento).

É de referir que ambas as unidades curriculares se orientam por uma perspetiva de avaliação formativa, uma vez que não se baseiam apenas num elemento de mensuração de conhecimentos (o exame, que irá ser alvo de estudo), mas também, em ambas, num trabalho prático, executado em grupo e acompanhado, com tema relacionado com os conteúdos abordados nas unidades curriculares e que suscita a aplicação de metodologias nelas veiculadas. Para além disso, é proporcionado aos estudantes a oportunidade de contacto com o tipo de instrumento de avaliação de conhecimentos, em aula preparatória da avaliação escrita, sendo discutidas com os alunos algumas questões de exame do ano anterior e respetivas respostas, com o intuito, não só de preparar a avaliação, mas também de ajudar os estudantes a sedimentar e organizar conhecimentos e a dar uma orientação ao estudo. Também é de referir que no decorrer do semestre, as docentes destas unidades curriculares, proporcionam semanalmente espaço para apoio tutorial individual ou em pequeno grupo, esclarecendo dúvidas e promovendo um estudo orientado, frequente e sistematizado. Ainda, cada estudante, quer reprove numa época de exame, quer nem sequer reprove, tem oportunidade de consultar a correção do seu exame, e receber, da parte das docentes, *feedback* pessoal sobre a qualidade das suas respostas, bem como do fundamento das suas classificações.

Desta forma, procura-se ultrapassar a avaliação sumativa, de mera certificação das aprendizagens, enquadrada na "geração da medida" (Guba & Lincoln, 1989), uma vez que os exames constituem instrumentos de avaliação de conhecimentos e de certificação de competências adquiridas, mas servem também finalidades formativas, orientando e dando aos estudantes um papel ativo em todo processo de ensino-aprendizagem. Posto isto, a presente investigação, apesar de se situar, em sentido estrito, no enquadramento teórico da docimologia clássica ou "negativa", acrescenta uma dimensão que vai para além da avaliação quantitativa ou instrumental: pretende retomar a aplicação de técnicas de estudo docimológico ao aperfeiçoamento de instrumentos de avaliação, a utilizar quer numa perspetiva de avaliação sumativa, quer numa perspetiva de avaliação formativa.

4.2. OBJETIVOS DO ESTUDO

A revisão de literatura efetuada revelou que a análise de instrumentos de avaliação de conhecimentos emerge como uma temática essencial a ser explorada, no âmbito da Psicologia Educacional, não só pela sua pertinência, como também por ser tema que tem vindo a ser negligenciado, na investigação e na literatura. A Psicologia Educacional abrange diversas áreas e conteúdos, mas a promoção da melhoria da qualidade do ensino e da educação é transversal a todo o seu domínio. A avaliação, enquanto instrumento de mensuração de aprendizagens, é um elemento incontornável no percurso escolar de todos os alunos, e assume um papel que, por vezes, define o seu posterior sucesso académico. Nas universidades, esta avaliação é configurada pelo professor, cabendo a este identificar a forma mais adequada para avaliar e mensurar as aprendizagens realizadas pelos estudantes. Paralelamente, deverá caber-lhe também a responsabilidade de averiguar a qualidade dos instrumentos de avaliação que constrói e utiliza para atribuir classificações aos seus estudantes.

Assim, o presente estudo é de carácter exploratório, relativamente aos dados mas também à própria metodologia, pelo que não parte de um conjunto de hipóteses, mas antes pretende proceder à avaliação docimológica das provas de exame de duas unidades curriculares obrigatórias na formação de psicólogos, na Faculdade de Psicologia da Universidade de Lisboa, Psicometria e Psi. Diferencial. Intenta-se realizar uma análise crítica dos instrumentos de avaliação de conhecimentos em uso, tomando por referência os objetivos da formação, bem como fundamentar opções futuras de revisão e aperfeiçoamento do formato e da escolha e organização dos conteúdos das provas de exame.

Como tal, este estudo visa:

– Analisar e avaliar dois formatos distintos de avaliação de conhecimentos aplicados na mesma unidade curricular, Psi. Diferencial, em dois anos letivos consecutivos, 2010/11 e 2011/12, nas três épocas de exame (1ªépoca, 2ªépoca e Época Especial e Específica);

– Analisar e avaliar o mesmo formato de avaliação aplicado em duas unidades curriculares, Psicometria e Psi. Diferencial, no mesmo ano letivo 2011/12, nas três épocas de exame.

– Estabelecer comparações entre resultados obtidos em dois anos letivos, para a mesma unidade curricular (Psi. Diferencial) e entre resultados das duas unidades curriculares, no mesmo ano letivo (2011/12), tendo em conta as épocas de exame.

– Explorar os percursos dos estudantes repetentes, analisando a evolução das classificações, ao longo das épocas de exame.

– Estabelecer comparações entre os exames de Psi. Diferencial nos dois formatos de avaliação de conhecimentos, em anos letivos diferentes, 2010/11 e 2011/12, com recurso à amostra dos estudantes repetentes.

Para alcançar estes objetivos, será efetuado um estudo metrológico dos instrumentos das duas unidades curriculares, nos dois anos letivos e nas três épocas de exames, bem como a análise comparativa de três formatos de itens (escolha múltipla, escolha múltipla com justificação e desenvolvimento) quanto ao valor discriminativo dos itens (correlação com a classificação final na exame).

III. METODOLOGIA

1. CARACTERIZAÇÃO DA AMOSTRA

A amostra é composta por 925 participantes, alunos do Mestrado Integrado em Psicologia, na Faculdade de Psicologia (FP), da Universidade de Lisboa (UL), estando a frequentar as duas unidades curriculares sob análise: Psi. Diferencial e Psicometria, nos anos letivos de 2010/11 e 2011/12. Uma vez que se procedeu à recolha dos dados após a realização das respetivas avaliações de conhecimentos, não foi possível obter informações adicionais quanto aos dados pessoais de cada participante.

Como se pode observar no Quadro 1, 785 participantes são do sexo feminino, constituindo cerca de 85% da amostra, e 140 participantes do sexo masculino, cerca de 15% da amostra (o que corresponde aproximadamente à proporção de alunos dos dois géneros que frequentam a FP da UL), perfazendo um total de 925 participantes. Especificamente, cerca de 321 participantes (correspondendo a 35% da amostra) realizaram o exame de Psicometria em 2011/12, e 604 participantes (65%) realizaram o exame de Psi. Diferencial, em cada um de dois anos letivos distintos: em 2010/11 foram 267 participantes (cerca de 29% da amostra) e em 2011/12, 337 participantes (cerca de 36% da amostra). A amostra encontra-se subdividida em função das três épocas de exame, sendo que, em ambos os anos letivos, e nas duas unidades curriculares, estiveram presentes, na 1ºÉpoca de exame 423 participantes (cerca de 46%), na 2ªÉpoca 346 participantes (cerca de 37%), e na Época Especial e Específica, 156 participantes (cerca de 17%).

Quadro 1 - Caracterização da Amostra

Sexo		Ano Letivo	Unidade Curricular	Época de exame			Total
				1ªÉpoca	2ªÉpoca	Época Especial	
M		2010/11	Psi. Diferencial	14	16	6	36
		2011/12	Psicometria	20	21	10	51
			Psi. Diferencial	23	20	10	53
	Totais	2011/12	Psicometria + Psi. Diferencial	43	41	20	104
		2010/11 + 2011/12	Psi. Diferencial	37	36	16	89
		2010/11 + 2011/12	Psicometria + Psi. Diferencial	57	57	26	140
F		2010/11	Psi.Diferencial	79	102	50	231
		2011/12	Psicometria	131	95	44	270
			Psi. Diferencial	156	92	36	284
	Totais	2011/12	Psicometria + Psi. Diferencial	287	187	80	554
		2010/11 + 2011/12	Psi. Diferencial	235	194	86	515
		2010/11 + 2011/12	Psicometria + Psi. Diferencial	366	289	130	785
Total – F+M		2010/11	Psi. Diferencial	93	118	56	267
		2011/12	Psicometria	151	116	54	321
			Psi. Diferencial	179	112	46	337
		2010/11 + 2011/12	Psi. Diferencial	272	230	102	604
		2010/11 + 2011/12	Psicometria + Psi. Diferencial	423	346	156	925

2. DESCRIÇÃO DOS INSTRUMENTOS

Os instrumentos alvo desta investigação são exames finais escritos de avaliação de conhecimentos das unidades curriculares de Psi. Diferencial e Psicometria, realizados no âmbito do Mestrado Integrado em Psicologia, em diferentes anos letivos. Em causa estão dois tipos de exames, com formatos distintos, mas ambos constituídos por perguntas de escolha múltipla com justificação (resposta breve) e de desenvolvimento.

No ano letivo de 2010/11, o exame de Psi. Diferencial era composto por duas partes, uma avaliando conhecimentos da teoria psicométrica (Teoria da Medida e Teoria dos Testes) e outra, aspetos da prática psicométrica (metodologia, técnicas e práticas de utilização, questões deontológicas, etc.) – parte Teórica e parte Prática, respetivamente. Em ambas as partes do exame, eram apresentadas cinco perguntas de escolha múltipla, com pedido de justificação (resposta curta de cinco linhas), e uma pergunta de desenvolvimento (resposta longa de cerca de duas páginas). Para facilitar os cálculos das classificações, cada parte do exame foi cotada para 20 valores (as perguntas de escolha múltipla e a pergunta de desenvolvimento equivaliam a 10 valores cada, em ambas as partes), sendo a nota final a média das duas partes. Para cada item de escolha múltipla, o acerto na alternativa correta valeria uma pontuação de 0.5 valores, sendo que a ponderação dos restantes valores (1.5 valores) era considerada a partir da qualidade da justificação dada à opção de resposta (podendo a cotação no item variar entre 1.0, 1.5 e 2.0). Para a apreciação final, na unidade curricular, o exame tinha uma ponderação de 80%, ou seja, de 16 valores em 20, sendo que os restantes 4 valores diziam respeito a um trabalho de grupo realizado ao longo do semestre.

No ano letivo de 2011/12, os exames de Psi. Diferencial e de Psicometria passaram a ser constituídos por dez perguntas de escolha múltipla, que tanto podiam ser teóricas, como práticas como ainda de articulação teórico-prática[1], com justificação breve, e uma pergunta de desenvolvimento, avaliando também conhecimentos teóricos, práticos e teórico-práticos. Cada item de escolha múltipla tem uma pontuação máxima de 1.0 valor, sendo que o acerto na alternativa correta equivale a 0.25 valores, e a ponderação dos valores restantes (0.75 valores) era atribuída em função da qualidade da justificação dada à opção de resposta (podendo variar entre 0.5, 0.75 e 1.0). Paralelamente, para a apreciação final os exames passaram a ter uma ponderação de 70%, ou seja, de 14 valores, em Psicometria, e de 80%, 16 valores, em Psi. Diferencial, em função da natureza dos trabalhos práticos de cada uma das unidades curriculares. Em ambas, os alunos realizaram um trabalho prático, em grupo, mas de natureza distinta, com um peso na nota final, de 6 valores em Psicometria e 4 valores em Psi. Diferencial.

É de salientar que a justificação dada a cada item de escolha múltipla é avaliada a partir de critérios bastante específicos, determinados *a priori* pelas docentes. Paralelamente, para a cotação da pergunta de desenvolvimento os critérios de classificação, são previamente definidos e partilhados por ambas as docentes na cotação

[1]A partir deste ano, o exame deixou de se dividir em partes Teórica e Prática por se ter abandonado a exigência de nota ≥ 9.5 valores em cada parte para aprovação. A parte de escolha múltipla do exame é sempre composta por 5 perguntas teóricas, 3 práticas e 2 teórico-práticas. O menor peso dado à parte prática do exame deve-se ao facto de esta ser sobretudo avaliada através da realização do trabalho prático.

do item, com o intuito de reduzir as oscilações interjuízes. De modo a facilitar a cotação da pergunta de desenvolvimento e das justificações e a comparação entre exames, aquando da sua classificação também é realizado um comentário qualitativo às respostas dos estudantes. Recorde-se que os critérios de classificação das respostas abertas são genericamente apresentados aos alunos, na aula de preparação para exame, juntamente com exemplos de itens do ano anterior, de modo a que compreendam os objetivos a atingir em função do tema de cada questão.

Em todos os elementos avaliativos descritos, era necessário, para a aprovação nas unidades curriculares, uma classificação de exame igual ou superior a 9,5 valores, sendo que, no ano letivo 2010/11, a aprovação na unidade curricular (Psi. Diferencial) era dependente da aprovação (valor igual ou superior a 9,5 valores) em ambas as partes Teórica e Prática, do exame. No ano seguinte, deixou de ser exigida a classificação de 9,5 valores em cada parte do exame (neste caso escolha múltipla e desenvolvimento), pois abandonou-se a estrutura separada entre partes Teórica e Prática, e passou a considerar-se apenas a exigência de 9,5 valores como classificação mínima do conjunto do exame. Também é de salientar que em todos os exames de avaliação de aprendizagens, nas unidades curriculares sob análise, o tempo estabelecido para completar a prova era de 2 horas e 30 minutos.

Uma vez que temos dois formatos de exame diferentes, com cotações distintas em cada item, ao preparar os dados para o presente estudo houve necessidade de realizar uma conversão das cotações de cada item, para se poder proceder à comparação entre exames. Desta forma, no exame de Psi. Diferencial, em 2010/11, os itens de escolha múltipla foram recodificados de modo a equivalerem ao mesmo tipo de itens dos restantes exames, expressando-se numa mesma escala de classificação (0.25 valores para opção correta, mesmo que não justificada ou mal justificada e 0.5, 0.75 ou 1.0 em função da qualidade da justificação).

3. Procedimento de Recolha de Dados

Os dados para o presente estudo foram recolhidos aquando da realização das avaliações académicas, nas duas unidades curriculares anteriormente descritas, nos anos letivos 2010/11, no 2º semestre, e em 2011/12 nos 1º e 2º semestre, em três distintas épocas de exames (o que perfaz 9 exames diferentes, 3 exames x 3 épocas de exame). Visto que no momento da realização destas avaliações académicas não estava prevista a

realização deste estudo, não se procedeu na altura à solicitação de consentimentos informados para esta utilização dos dados. E por ser difícil a localização de todos os participantes para tal propósito, adotou-se em alternativa uma utilização totalmente anónima dos dados. Isto implicou que a investigadora, ao longo de todo o desenvolvimento desta monografia, não tivesse qualquer acesso ao conteúdo das respostas de exame (caligrafia dos estudantes), nem a qualquer outra forma de contacto com a identidade dos estudantes que os realizaram. Para garantir este anonimato dos participantes, os dados para a presente investigação foram coligidos e fornecidos pela docente de ambas as unidades curriculares com o apoio do respetivo monitor[2] através de uma Ficha de Classificação de Exame (FCE), para cada participante avaliado em cada época de exame. Assim sendo, estas fichas, previamente construídas pela docente (ver modelo de exame e FCE no Anexo 1), contêm as alternativas de respostas escolhidas pelos estudantes, nos itens de escolha múltipla, a classificação atribuída em função da justificação das respostas e a classificação atribuída na pergunta de desenvolvimento, bem como um comentário qualitativo às justificações das respostas de escolha múltipla e de desenvolvimento. Foi omitido desta ficha qualquer dado identificativo do estudante participante.

Ainda com o intuito de salvaguardar o anonimato dos dados, cada estudante foi identificado por um número convencional de participante, atribuído pela docente, através de uma transformação monótona operada sobre o número do aluno da FP, cujo procedimento de cálculo nunca foi revelado à investigadora, com o intuito de que o mesmo aluno, em diferentes momentos avaliativos, recebesse o mesmo número de participante, possibilitando o emparelhamento de dados dos mesmos alunos, obtidos em diferentes exames. Isto tornou possível um estudo longitudinal dos estudantes que repetiram mais do que uma vez o exame na mesma unidade curricular. Para além desta identificação convencional, apenas foi retida a informação do género do estudante.

4. METODOLOGIAS UTILIZADAS PARA A ANÁLISE DE ITENS DE ESCOLHA MÚLTIPLA.

As instituições educacionais utilizam uma larga variedade de instrumentos de avaliação para mensurar os seus estudantes, desde o uso de perguntas de escolha múltipla, de resposta breve/curta, de resposta de desenvolvimento, de resolução de problemas a apresentações. Segundo alguns autores, as perguntas de escolha múltipla

[2] Pelo investimento direto e muito moroso na preparação dos dados para este estudo, agradece-se à docente das unidades curriculares, a Prof.ª Dr.ª Maria João Afonso e ao respetivo Monitor, Dr.º Tiago Cabaço, por tornarem possível esta investigação, preservando o anonimato dos estudantes, sem contudo perder a sua identificação necessária para o emparelhamento dos dados relativos a cada estudante (por exemplo, no caso dos estudantes repetentes).

são utilizadas em larga escala, nas universidades, em detrimento das perguntas de desenvolvimento, dadas as suas numerosas vantagens (Bacon, 2003; DiBattista & Kurzawa, 2011). Porém, são escassos os estudos que se debruçam sobre os formatos avaliativos, que são utilizados em sala de aula, levando a que em geral os testes/exames utilizados em contexto escolar tenham qualidades psicométricas desconhecidas.

Para a análise dos resultados é, contudo, importante perceber que critérios devem ser explorados, para averiguar a qualidade dos itens e a escassez de literatura que possa servir de fundamento a esta definição de critérios de avaliação dos itens é talvez responsável pela escassez de estudos desta natureza. As perguntas de escolha múltipla são um elemento significativo nos formatos avaliativos anteriormente apresentados e, para averiguar a sua qualidade é pertinente explorar três características chaves: a dificuldade dos itens, o seu poder discriminativo e a eficiência dos distratores (alternativas erradas).

Segundo DiBattista e Kurzawa (2011), o índice de dificuldade dos itens diz respeito à proporção de respostas certas selecionadas pelos examinados, podendo variar entre 0 (ninguém selecionou a resposta correta) e 1 (todos selecionaram a resposta correta). Assim, um valor elevado representa um item mais fácil, enquanto um valor baixo representa um item difícil (Ebel & Frisbie, 1986). Kline (2005) propõe, para a medição psicológica, que o índice de dificuldade deverá encontrar-se dentro do intervalo de .20 a .80, mas outros autores como Ebel e Frisbie (1986) determinam um intervalo de .30 a .90. Scialfa, Legare, Wenger e Dingley (2001) apontam para um valor ótimo de índice de dificuldade de .50, por seu lado, Colbert (2001) indica para um item de escolha múltipla com três alternativas de resposta, um valor ideal de 0.665.

Outro índice determinante na qualidade dos itens de escolha múltipla, é o seu poder discriminativo, isto é, a capacidade de um item discriminar entre níveis altos e baixos, no que está a ser avaliado, portanto, de medir eficazmente as diferenças individuais entre os examinados (Haladyna, 2004). Por outras palavras, este índice estima o poder preditivo de um item em relação ao desempenho final no teste (Colbert, 2001). Os valores deste índice variam entre -1.0 e 1.0, mas Kline (2005) estabelece que para um item discriminar de forma efetiva, deverá apresentar um mínimo valor de .30, embora considere de .20 para cima valores aceitáveis. Assim, um valor positivo, indica que, tal como seria desejável, os examinados que obtiveram um melhor desempenho no teste escolheram mais vezes a opção correta nesse item do que os que tiveram um pior

desempenho, ou seja, que o item só por si se mostrou capaz de discriminar as diferenças individuais na variável que está a medir. Um valor negativo indica exatamente o inverso, ou seja, os examinados que tiveram um pior desempenho no teste, selecionaram a resposta correta mais vezes, do que os obtiveram um melhor desempenho (Kline, 2005); este constitui um padrão atípico ou distorcido de resposta e identifica um item não discriminativo. Este índice está relacionado com a dificuldade do item (proporção de acertos), uma vez que itens muito fáceis (proporção de acertos >.80) ou muito difíceis (<.20) não discriminam de forma eficiente entre alunos com melhores e piores desempenhos (Ebel & Frisbie, 1986). Apesar de ser escassa a literatura existente sobre a qualidade dos itens utilizados na avaliação das aprendizagens, a disponível indica-nos que a média do coeficiente de discriminação, para um teste ser eficiente, deverá ser superior a .20 (DiBattista & Kurzawa, 2011).

O poder discriminativo de um item depende também da qualidade dos distratores (das alternativas incorretas), uma vez que um distrator bem construído deverá parecer menos plausível para os examinados com mais conhecimentos, do que para os restantes. Um distrator eficiente é selecionado por pelo menos 5% dos examinados (DiBattista & Kurzawa, 2011; Colbert, 2001; Kline, 2005) – caso contrário será evidentemente errado mesmo para os estudantes com poucos conhecimentos; e é selecionado mais vezes pelos examinados com piores classificações do que pelos examinados com melhores classificações (DiBattista e Kurzawa, 2011).

Na presente investigação, procedeu-se também a uma análise dos itens de escolha múltipla, quando dicotómicos, isto é, quando o examinado acertou na alternativa correta, sem tomar em consideração a parte da classificação decorrente do grau de exatidão da sua justificação: o item foi cotado como 1 se correto e como 0 se o examinado não selecionou a alternativa correta, ou se não selecionou nenhuma das opções de resposta. Esta análise teve como objetivo perceber se existe uma diferença na qualidade metrológica dos itens de escolha múltipla, quando se toma em consideração a justificação, ou quando apenas se considera o acerto ou a falha na seleção da resposta correta, evitando a tarefa de classificação da qualidade das justificações.

O tratamento estatístico dos dados foi realizado mediante a utilização do *Software Statistical Package for Social Sciences (SPSS) 20.0 for Windows*.

IV. ANÁLISE DE RESULTADOS

O Quadro 2, resume as estatísticas descritivas de cada exame analisado, apresentando as médias e desvios-padrão dos diferentes itens (escolha múltipla e desenvolvimento), bem como a amplitude das classificações. Podemos, desde logo, verificar que os dois formatos de exame produziram resultados diferentes.

Nos formatos avaliativos que dizem respeito à unidade curricular de Psi. Diferencial, no ano letivo de 2010/11, que são divididos em duas partes (Teórica e Prática), as médias das classificações finais são mais baixas quando comparadas com o formato dos outros exames. Neste exame, podemos também perceber que as partes Teórica e Prática apresentam médias distintas, parecendo que a parte Teórica foi ligeiramente mais fácil do que a parte Prática. Interessante é também verificar que o item de desenvolvimento obteve médias superiores, comparativamente aos itens de escolha múltipla[3]. Quanto ao exame de Psi. Diferencial com outro formato, já descrito anteriormente, realizado no ano letivo de 2011/12, as médias das classificações finais são superiores às observadas no exame da mesma unidade curricular, realizado no ano letivo anterior, sendo que os itens de desenvolvimento obtiveram também médias superiores aos itens de escolha múltipla. Porém, verifica-se um aumento nas médias e uma maior variabilidade nos resultados dos itens de escolha múltipla, sendo este um indicador de que este exame é mais capaz de discriminar diferenças interindividuais nos conhecimentos dos estudantes. No exame de época especial, denota-se uma descida na média de classificação final do exame, podendo significar um maior grau de dificuldade do exame, embora também um menor nível de conhecimento por parte dos estudantes que recorrem a essa época. No exame de Psicometria, do ano letivo 2011/12, as diferenças entre as médias dos itens de escolha múltipla e de desenvolvimento não são tão díspares, verificando-se até, na época especial, uma superioridade na média dos itens de escolha múltipla.

[3] É de assinalar que, anteriormente ao ensaio destes novos formatos de exame iniciado em 2010/11, por cerca de 30 anos, os exames escritos destas unidades curriculares eram constituídos exclusivamente por dois temas para desenvolvimento, um incidindo na matéria teórica e outro na matéria prática. Pretendeu-se, desde 2010/11, verificar a adequação de outros formatos de itens para a avaliação dos conhecimentos adquiridos nestas unidades curriculares e é precisamente neste âmbito que se insere o presente estudo

Quadro 2 - Médias e desvio-padrão ($\bar{x}$ (dp)), amplitude dos resultados totais dos dois tipos de questões de exame: escolha múltipla e desenvolvimento (itens não dicotómicos)

		Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ªÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ªÉpoca (n=116)	Ép.Esp. (n=54)
Escolha Múltipla (0 – *5)*	Teórica (Psi.Diferencial 2010/11)	1.58 (.93)	1.57 (.95)	2.08 (.92)						
Amplitude		*0 - 4.00*	*0 - 4.00*	*.50 - 3.75*						
Desenvolvimento (0 – *5)*		2. 52 (1.25)	2.80 (1.06)	2.16 (.98)						
Amplitude		*0 – 4.75*	*0 –4.75*	*0 – 4.25*						
Escolha Múltipla + Desenvolvimento (0 – 10)		4.10 (1.93)	4.37 (1.77)	4.23 (1.67)						
Amplitude		0 - 8.25	.25 - 8.50	1.0 - 8.00						
Escolha Múltipla (0 – *5)*	Prática (Psi. Diferencial 2010/11)	1.31 (.81)	2.05 (1.15)	1.50 (.81)						
Amplitude		*0 – 3.75*	*0 – 4.50*	*0 – 3.50*						
Desenvolvimento (0 – *5)*		2.13 (1.18)	2.42 (.96)	2.41 (.86)						
Amplitude		*0 – 4.00*	*0 – 4.00*	*0 – 3.75*						
Escolha Múltipla + Desenvolvimento (0 – 10)		3.44 (1.69)	4.47 (1.84)	3.92 (1.43)						
Amplitude		*0 – 7.00*	*0 – 7.75*	*.25 - 6.25*						
Escolha Múltipla (0 – 10)	Teórica + Prática	2.89 (1.39)	3.62 (1.77)	3.58 (1.36)						
Amplitude		*.50 – 6.75*	*0 – 8.50*	*1 – 6.50*						
Desenvolvimento (0 – 10)		4.65 (2.23)	5.22 (1.87)	4.57 (1.64)						
Amplitude		*0 – 8.50*	*0 – 8.50*	*0 – 7.50*						
Escolha Múltipla + Desenvolvimento (0 – 10)		7.54 (3.32)	8.84 (3.30)	8.15 (2.68)						
Amplitude		*.50- 14.75*	*.75 -16.25*	*2.5 - 13.00*						
Escolha Múltipla (0 – 10)	Total				4.35 (2.17)	3.89 (1.89)	3.83 (2.12)	4.89 (2.20)	3.98 (1.82)	5.40 (2.01)
Amplitude					*0 - 10.0*	*.50 - 8.50*	*.50 - 8.0*	*.25– 9.25*	*0 - 8.75*	*1.0 - 9.0*
Desenvolvimento (0 – 10)					5.49 (2.03)	5.26 (1.83)	4.01 (1.89)	4.96 (2.14)	4.04 (2.30)	4.84 (1.74)
Amplitude					*0 - 9.0*	*0 – 8.50*	*0 - 8.0*	*0 - 9.0*	*0 - 9.5*	*0 - 8.0*
Escolha Múltipla + Desenvolvimento (0 – 10)					9.84 (3.72)	9.15 (3.21)	7.84 (3.73)	9.85 (3.99)	8.01 (3.67)	10.24 (3.25)
Amplitude					*1.0 -18.75*	*1.5 - 16.5*	*.75 - 16.0*	*1.5 – 17.5*	*.50 -17.50*	*1.0 -16.75*

No Quadro 3, estão apresentados os resultados, apenas relativos à escolha múltipla, dos itens cotados de forma dicotómica, uma vez que se pretende analisar e comparar o funcionamento dos itens de escolha múltipla em ambos os formatos, dicotómico e não

dicotómico. As conclusões a retirar deste quadro são idênticas às mencionadas anteriormente, naturalmente, sendo de maior interesse a comparação que adiante se fará, relativa à qualidade psicométrica dos itens.

Quadro 3 - Médias e desvio-padrão ($\bar{x}$ (dp)), mínimos e máximos dos resultados totais dos itens de escolha múltipla (itens dicotómicos)

		Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép. Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép. Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép. Esp. (n=54)
Escolha Múltipla (0 – *5)*	Teórica	2.73 (1.08)	2.36 (1.11)	3.04 (1.08)						
Amplitude		*0 – 5*	*0 – 5*	*1 – 5*						
Escolha Múltipla (0 – *5)*	Prática	2.04 (.92)	2.65 (1.36)	2.14 (1.00)						
Amplitude		*0 – 4*	*0 – 5*	*0 – 5*						
Escolha Múltipla (0 – 10)	Teórica + Prática	4.77 (1.42)	5.01 (1.99)	5.18 (1.53)						
Amplitude		*2 – 8*	*0 – 10*	*2 – 8*						
Escolha Múltipla (0 – 10)	Total				5.72 (2.08)	4.97 (1.79)	5.48 (1.95)	6.29 (1.87)	5.46 (1.83)	6.48 (1.75)
Amplitude					*0 – 10*	*1 – 9*	*2 – 9*	*1 – 10*	*0 – 10*	*3 -10*

No Quadro 4, são apresentados os índices de dificuldade dos itens de escolha múltipla não dicotómicos, que poderão variar de 0 a 1.

Verificamos que no conjunto dos nove exames, num total de 90 itens, apenas 10 itens se encontram fora do intervalo assinalado. O item 4, referente ao exame de Psi. Diferencial, 2010/11, na 2ªépoca, e o item 5, do exame de Psicometria, 2011/12, da época especial, apresentam valores de .11 e .13 respetivamente, espelhando uma maior dificuldade dos examinados em selecionar a alternativa correta. Os itens restantes representam perguntas mais fáceis, onde a maioria dos examinados selecionou a resposta correta, variando de .81 a .96 . Constatamos que os itens de escolha múltipla na parte Teórica são sensivelmente mais fáceis que os da parte Prática, nos exames de Psi. Diferencial, 2010/11, à exceção do exame de 2ª época. Tendo em conta a média da proporção de respostas corretas (média do índice de dificuldade) do total de cada exame, observamos valores entre .48 e .65, significando que em média os itens têm uma dificuldade adequada, uma vez que se aproxima de .50, como seria ideal.

Quadro 4 - Índice de dificuldade dos itens de escolha múltipla (itens não dicotómicos)

		Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
Item	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Èp. Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép. Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép. Esp. (n=54)
1	Teórica (Psi.Diferencial 2010/11)	**.59**	**.21**	**.41**	**.60**	**.30**	**.61**	.85	**.32**	.83
2		**.52**	**.79**	**.64**	**.51**	**.56**	**.74**	.81	**.20**	.96
3		**.66**	**.69**	**.70**	**.53**	**.31**	**.59**	.85	**.48**	**.52**
4		**.63**	.11	**.64**	**.80**	**.74**	**.63**	**.65**	**.79**	**.43**
5		**.33**	**.57**	**.64**	**.77**	**.63**	**.52**	**.25**	**.80**	.13
6	Prática (Psi. Diferencial 2010/11)	**.29**	**.58**	**.64**	**.54**	**.63**	**.44**	**.47**	**.42**	**.59**
7		**.36**	**.54**	**.23**	**.72**	**.36**	**.48**	**.55**	.83	**.57**
8		**.70**	**.76**	**.59**	**.54**	**.50**	**.70**	**.56**	**.55**	.82
9		**.44**	**.26**	**.20**	**.23**	**.28**	**.35**	**.53**	**.53**	**.69**
10		**.25**	**.51**	**.48**	**.48**	**.68**	**.44**	**.78**	**.53**	.94
Média da proporção de Respostas Corretas	Teórica	**.55**	**.47**	**.61**	**-**	**-**	**-**	**-**	**-**	**-**
	Prática	**.41**	**.53**	**.43**	**-**	**-**	**-**	**-**	**-**	**-**
	Total	**.48**	**.50**	**.52**	**.57**	**.50**	**.55**	**.63**	**.55**	**.65**

Valores a negrito dentro do intervalo adequado [.20 - .80] (Kline, 2005)

Outro indicador importante é o índice de discriminação dos itens, que se encontra apresentado no Quadro 5, representando a correlação item-total, isto é, o poder preditivo do item, em relação ao desempenho total no exame.

Quadro 5 - Índice de discriminação dos itens de escolha múltipla (itens não dicotómicos)

		Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
Item	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
1	Teórica (Psi. Diferencial 2010/11)	.18	.13	-.04	**.29**	.16	**.45**	**.62**	**.32**	**.44**
2		**.21**	**.41**	.17	**.32**	**.36**	**.44**	**.41**	.17	**.30**
3		.16	.18	.16	**.45**	.18	**.34**	**.41**	**.38**	**.36**
4		**.30**	**.31**	.03	**.41**	**.26**	**.29**	**.22**	**.33**	**.30**
5		.18	**.33**	.17	**.45**	**.35**	**.48**	.08	**.36**	**.20**
6	Prática (Psi. Diferencial 2010/11)	**.21**	**.27**	.03	**.51**	**.42**	**.50**	**.45**	.19	**.29**
7		.12	**.37**	**.28**	**.41**	**.21**	.13	**.47**	**.47**	**.52**
8		.16	**.33**	.04	**.45**	**.41**	**.50**	**.50**	**.36**	**.38**
9		.01	**.28**	-.08	.19	.13	**.29**	**.50**	**.31**	**.44**
10		.00	**.30**	-.01	**.43**	**.41**	**.45**	**.41**	.14	**.42**
Amplitude		.00 -.30	.13 - .41	-.08 -.28	.19 - .51	.13 - .42	.13 -.50	.08 - .62	.14 - .47	.20 - 52
Média do Índice de Discriminação	Teórica	**.21**	**.27**	.10	-	-	-	-	-	-
	Prática	.10	**.31**	.05	-	-	-	-	-	-
	Total	.15	**.29**	.07	**.39**	**.29**	**.39**	**.41**	**.30**	**.36**

Valores significativos a negrito, indicadores de um bom (>.30) ou de um aceitável (> .20) poder discriminativo dos itens (Kline, 2005; DiBattista & Kurzawa, 2011)

Como podemos verificar, existe uma grande diferença no índice de discriminação, entre os dois formatos. No primeiro formato de exame (Psi. Diferencial, 2010/11), apenas na 2ª época temos vários itens que correspondem aos critérios selecionados, significando que os itens não estão a discriminar eficientemente entre os examinados que obtiveram um bom desempenho dos que tiveram um pior desempenho. Os itens 1, 9 e 10, do exame de Psi. Diferencial, 2010/11, da época especial, apresentam valores negativos, mas pertos de zero, refletindo um fraco poder discriminativo. Nos restantes exames, com um formato diferente, verificamos ser maior a quantidade de itens a discriminar entre examinados de forma eficaz. Quanto à média do índice de discriminação (média das correlações item-total), verificamos que todos os itens são indicadores de um bom poder discriminativo, variando entre .29 e .41, à exceção dos exames de 1ª época e época especial, de Psi. Diferencial, 2010/11, onde são apresentados índices de .15 e .07, respetivamente, indicando um fraco poder discriminativo. De notar que estes foram os primeiros exames elaborados pelas docentes

neste novo formato, incluindo questões de escolha múltipla, o que pode justificar em parte os resultados inferiores da sua qualidade metrológica.

Torna-se importante perceber, agora, o índice de discriminação dos itens de escolha múltipla quando dicotómicos (Quadro 6), com o intuito de compreender o seu poder discriminativo quando não se considera a justificação dada pelo estudante, mas apenas a seleção da resposta certa.

Quadro 6 - Índice de discriminação dos itens de escolha múltipla (itens dicotómicos)

		Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
Item	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
1	Teórica Psi. Diferencial 2010/11)	.06	.03	-.01	.08	.03	.14	**.22**	.15	.15
2		-.03	.16	-.00	.17	.16	**.25**	**.20**	.08	.06
3		-.17	.13	.12	**.26**	.04	.05	.12	**.24**	**.32**
4		-.02	**.24**	-.12	**.25**	.07	-.01	.14	.14	**.24**
5		.12	**.26**	.03	**.22**	.19	**.33**	-.07	.11	.05
6	Prática (Psi. Diferencial 2010/11)	-.07	**.24**	-.12	**.34**	**.26**	**.34**	**.37**	.12	.13
7		-.17	**.34**	**.28**	**.28**	.15	-.02	**.28**	.19	**.24**
8		-.03	**.27**	-.09	**.33**	.16	.18	**.23**	**.27**	**.29**
9		-.19	**.24**	-.11	.16	-.04	.05	**.26**	**.20**	**.33**
10		-.12	**.24**	-.07	**.25**	.18	**.37**	**.20**	.13	.13
Amplitude		-.19-.12	.03 - .27	-.12 -.28	.08 - .34	-.04 -.26	-.01-.37	-.07-.37	.08-.27	.06 - .32
Média do Índice de Discriminação	Teórica	-.01	.16	.00	-	-	-	-	-	-
	Prática	-.12	**.27**	-.02	-	-	-	-	-	-
	Total	-.06	**.22**	-.01	**.23**	.12	.17	**.20**	.16	.19

Valores significativos a negrito, indicadores de um bom (> .30) ou de um aceitável (> .20) poder discriminativo dos itens (Kline, 2005; DiBattista & Kurzawa, 2011)

Observamos uma significativa diferença nos resultados do índice de discriminação quando consideramos os itens de escolha múltipla na forma dicotómica. Temos vários itens com uma correlação item-total negativa, significando que não estão a discriminar entre os examinados com melhores e piores desempenhos. Sendo que nalguns itens (nomeadamente itens 3, 7 e 9, de Psi. Diferencial, 2010/11, 1ªépoca) os valores negativos traduzem que os examinados com piores desempenhos escolheram a

alternativa correta mais vezes que os restantes. Para além disso, a média do índice de discriminação nos diferentes exames, é menor, chegando a ser negativa, nos exames de 1ª época e de época especial, de Psi. Diferencial, 2010/11, e apenas 3 exames apresentam valores aceitáveis (.22, .23 e .20). Este é o primeiro indicador de que, nos itens de escolha múltipla, há vantagem em tomar em consideração a justificação da opção efetuada (resposta curta), pois tal coloca melhor em evidência as diferenças individuais nos conhecimentos teóricos e práticos de Psicologia Diferencial e de Psicometria.

No Quadro 7, encontra-se a análise da qualidade dos distratores, representando as proporções dos examinados que escolheram uma alternativa de resposta (A, B ou C) num item, sendo no quadro omitido a proporção de respostas na alternativa correta, em cada item.

Quadro 7 - Qualidade dos distratores dos itens de escolha múltipla: proporções dos examinados que selecionaram a alternativa de resposta (A, B ou C) no item.

			Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
Alternativas de Resposta		Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
Item 1	A	Teórica (Psi. Diferencial 2010/11)	.26	.35	-	.28	.36	.33	.05	.34	-
	B		.12	-	.13	.11	.32	-	.09	.35	.13
	C		-	.42	.41	-	-	.07	-	-	.04
Item 2	A		.17	.05	.23	.30	.28	.13	.15	.73	.02
	B		-	.14	.11	-	.12	.13	.04	-	.02
	C		.26	-	-	.19	-	-	-	.07	-
Item 3	A		.10	-	.07	.29	-	.24	.05	.22	.35
	B		.24	19	-	-	.32	-	-	-	-
	C		-	.12	.23	.18	.36	.13	.10	.27	.13
Item 4	A		-	.51	.18	.09	.11	.17	.11	.10	-
	B		.10	-	.16	.11	-	-	-	.10	.50
	C		.24	.36	-	-	.14	.20	.23	-	.07
Item 5	A		-	-	-	.11	.05	-	.16	-	.70
	B		.45	.19	.29	.11	-	.17	.58	.03	.17
	C		.22	.18	.09	-	.31	.30	-	.16	-
Item 6	A	Prática (Psi.Diferencial 2010/11)	.30	.26	-	.28	-	.15	.38	.13	-
	B		.40	.10	.20	-	.24	.41	.15	.45	.39
	C		-	-	.14	.18	.13	-	-	-	.02
Item 7	A		.16	.21	.46	.15	.04	.17	.19	-	.02
	B		-	-	-	.14	.60	-	-	.07	.41
	C		.41	.23	.27	-	-	.33	.25	.08	-
Item 8	A		.18	-	.20	-	.31	.17	.25	-	.13
	B		-	.07	-	.41	-	.13	.19	.20	.06
	C		.11	.15	.21	.04	.17	-	-	.25	-
Item 9	A		.17	.48	-	-	.28	-	-	.14	.13
	B		.30	.24	.39	.41	.42	.41	.20	.34	-
	C		-	-	.34	.36	-	.22	.25	-	.19
Item 10	A		-	.11	.30	.30	-	-	.13	.03	-
	B		.33	-	.19	.16	.14	.24	.09	-	.04
	C		.27	.35	-	-	.17	.30	-	.45	.02

Nota 1: A alternativa correta não está assinalada no quadro, apenas os distratores (alternativas incorretas).
Nota 2: Um distrator é eficiente se for selecionado por pelo menos 5% dos examinados (DiBattista & Kurzawa, 2011).

Verificamos no Quadro 7 que apenas 12 distratores (6%) são selecionados por menos de 5% dos examinados (<.05). No entanto, segundo DiBattista e Kurzawa

(2011), outro critério deverá ser verificado, de modo, a considerar os distratores eficientes: os examinados com piores desempenhos, deverão escolher os distratores mais vezes do que os examinados com melhores desempenhos. Essa análise foi realizada, constituindo dois grupos, sendo o Grupo 1 o dos examinados com as classificações mais baixas no conjunto dos itens de escolha múltipla, situadas no Quartil 1 (≤ Percentil 25) da distribuição das classificações, e o Grupo 2 o dos examinados com as classificações mais altas no conjunto dos itens de escolha múltipla, situadas no Quartil 4 (> Percentil 75). Verificou-se que os distratores (alternativas erradas) são, na esmagadora maioria dos itens, selecionados mais vezes pelo Grupo 1 (alunos com as classificações mais baixas na escolha múltipla) do que pelo Grupo 2 (alunos com as classificações mais altas) (cf Anexo 2 – Quadros 13, 14 e 15). Para além disso, é mais frequente um distrator não ser selecionado por ninguém no Grupo 2 no que pelo Grupo 1, e sempre que ninguém do Grupo 1 escolheu o distrator, também ninguém do Grupo 2 o selecionou.

Apenas três distratores são selecionados mais vezes pelo Grupo 2 do que pelo Grupo 1: a alternativa A do item 7, do exame de Psi. Diferencial (2011/12), da época especial; a alternativa B do item 5 do exame de Psicometria, de 1ª época; a alternativa A do item 5 do exame de Psicometria, de época especial. Também se verifica haver 5 distratores selecionados as mesmas vezes pelos dois grupos. Posto isto, podemos considerar que cerca de 89% dos distratores utilizados nestes exames foram eficientes, uma vez que respeitam os dois critérios propostos por DiBattista e Kurzawa (2011), talvez por terem sido construídos tomando em conta princípios cientificamente estabelecidos para a construção dos itens de escolha múltipla (Haladyna, 2004).

Ainda na perspetiva da análise dos itens, mas no âmbito do estudo da consistência interna das escalas constituídas por itens de escolha múltipla, mediante o cálculo do alfa de Cronbach, são apresentados, no Quadro 8, os coeficientes de alfa de Cronbach quando cada item é eliminado, que nos fornecem informação sobre a qualidade dos itens do ponto de visto do seu contributo para a consistência interna (Marôco & Garcia-Marques, 2006) bem como o alfa de Cronbach e o alfa de Cronbach estandardizado de cada escala, e por último a média das correlações inter-itens.

Quadro 8 - Consistência interna das partes de escolha múltipla (alfa de Cronbach), coeficientes alfa com cada item eliminado e correlações inter-itens - itens não dicotómicos

			Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
	Item	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
Alfa se item for omitido	1	Teórica (Psi. Diferencial 2010/11)	.37	.51	.29	.73	.62	.70	.69	.61	.66
	2		.35	.32	.11	.72	.58	.70	.72	.63	.69
	3		.39	.49	.11	.70	.62	.71	.72	.59	.68
	4		.28	.42	.25	.71	.61	.72	.75	.61	.69
	5		.37	.38	.11	.70	.59	.69	.76	.60	.70
	6	Prática (Psi. Diferencial 2010/11)	.05	.52	.11	.69	.57	.69	.72	.63	.69
	7		.17	.45	-.15	.71	.62	.74	.71	.57	.64
	8		.12	.48	.09	.70	.57	.69	.71	.60	.67
	9		.28	.51	.18	.74	.63	.72	.71	.61	.66
	10		.27	.50	.15	.71	.57	.70	.72	.65	.67
Alfa de Cronbach		Teórica	.41	.49	.22	-	-	-	-	-	-
		Prática	.23	.55	.10	-	-	-	-	-	-
		Total	.47	.65	.33	**.73**	.62	**.73**	**.74**	.64	**.70**
Alfa de Cronbach estandardizado		Teórica	.40	.50	.20	-	-	-	-	-	-
		Prática	.20	.55	.11	-	-	-	-	-	-
		Total	.43	.65	.30	**.73**	.61	**.72**	**.74**	.64	**.70**
Correlação Inter- Item		Teórica	.12	.17	.05	-	-	-	-	-	-
		Mínimo	.02	-.01	-.13	-	-	-	-	-	-
		Máximo	.25	.38	.33	-	-	-	-	-	-
		Prática	.05	.20	.02	-	-	-	-	-	-
		Mínimo	-.12	.11	-.13	-	-	-	-	-	-
		Máximo	.21	.28	.19	-	-	-	-	-	-
		Total	.07	.16	.04	.21	.14	.21	.22	.15	.19
		Mínimo	-.20	-.15	-.17	.00	-.10	-.08	-.11	-.10	-.07
		Máximo	.33	.38	.53	.41	.35	.52	.47	.43	.42

Valores a negrito, coeficientes de alfa de Cronbach aceitáveis, (≥.70), (Maroco & Garcia-Marques , 2006).
Nota: Os valores de alfa de Cronbach se o item for omitido, apresentados nos exames de Psi. Diferencial, 2010/11, dizem respeito às partes teóricas e prática separadamente e não ao total da escala.

Verificamos no Quadro 8, uma grande diferença nos índices entre os dois formatos de avaliação de conhecimentos apresentados. No primeiro formato (partes Teórica e Prática), encontramos, no exame de 1ª época e de época especial, alfas de Cronbach

relativos ao total da escala, abaixo do aceitável, com valores de .47 e .33, respetivamente. Observando os coeficientes de alfa de Cronbach para as partes Teórica e Prática, separadamente, dos dois exames acima mencionados, verificamos que a consistência interna destas escalas é muito baixa (entre .10 e .55), principalmente na parte Prática, sendo que alguns itens têm um grande peso na consistência interna da escala.

No outro formato avaliativo, constatamos coeficientes de alfa de Cronbach entre .64 e .74, demonstrando uma melhor consistência interna das escalas quando os itens de escolha múltipla não são divididos em parte teórica e prática. Quanto à correlação inter-itens, esta determina o grau em que cada item está relacionado com os restantes, dela dependendo os coeficientes alfa de Cronbach. Verificamos correlações médias inter-itens totais entre .04 e .22, que colocam em evidência o facto de em testes de conhecimentos ser necessário avaliar uma panóplia de conteúdos diversos, e consequentemente torna-se díifícil atingir correlações médias inter-itens elevadas (esta questão irá será explorada mais à frente neste trabalho).

No Quadro 9, são apresentados os mesmos indicadores, mas tendo em consideração os itens na forma dicotómica. Observamos, neste quadro, no primeiro formato avaliativo (partes Teórica e Prática) valores de alfa de Cronbach negativos, bem como a correlações inter-itens negativas ou próximas de zero (exames de 1ª época e época especial de Psi. Diferencial, 2010/11). Marôco e Garcia-Marques (2006) referem que um valor de alfa de Cronbach negativo poderá refletir um erro na codificação dos pontos dos itens, no entanto, no presente caso, na inspeção dos itens, não encontramos razões para tal suceder. O que nos leva a concluir que os itens de escolha múltipla, quando não se considera a justificação (a resposta curta), não estão a contribuir de forma eficaz para a consistência interna da escala, espelhando uma fraca contribuição para a avaliação de conhecimentos nestas unidades curriculares.

Quadro 9 - Consistência interna das partes de escolha múltipla (alfa de Cronbach), coeficientes alfa com cada item elimando e correlações inter-itens – itens dicotómicos

			Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
	Item	Partes	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
Alfa se item eliminado	1	Teórica	-.12	.38	.02	.56	.35	.40	.45	.40	.47
	2		-1.98$^{-.017}$	.28	.01	.53	.29	.36	.45	.42	.49
	3		.17	.31	-.16	.51	.34	.43	.47	.36	.41
	4		-.02	.23	.15	.51	.33	.45	.47	.40	.45
	5		-.21	.17	-.05	.52	.28	.32	.53	.41	.50
	6	Prática	-.37	.46	.05	.48	.24	.32	.38	.41	.49
	7		-.16	.38	-.57	.50	.29	.46	.42	.38	.44
	8		-.44	.43	.01	.48	.29	.38	.44	.35	.43
	9		-.13	.45	.01	.53	.37	.43	.43	.38	.41
	10		-.26	.45	-.02	.51	.28	.31	.45	.41	.48
Alfa de Cronbach		Teórica	-.03	.33	-.00	-	-	-	-	-	-
		Prática	-.35	.49	-.10	-	-	-	-	-	-
		Total	-.15	.54	.04	.54	.33	.42	.48	.42	.48
Alfa de Cronbach estandardizado		Teórica	-.03	.33	-.00	-	-	-	-	-	-
		Prática	-.34	.49	-.08	-	-	-	-	-	-
		Total	-.15	.53	.06	.54	.32	.42	.47	.41	.46
Correlação Inter-Item		Teórica	-.01	.09	.00	-	-	-	-	-	-
		Mínimo	-.13	-.09	-.21	-	-	-			
		Máximo	.16	.26	.24	-	-	-			
		Prática	-.05	.16	-.01	-	-	-	-	-	-
		Mínimo	-.12	.08	-.21	-	-	-			
		Máximo	.10	.25	.23	-	-	-			
		Total	-.01	.10	.01	.11	.05	.07	.08	.07	.08
		Mínimo	-.24	-.15	-.21	-.10	-.14	-.27	-.18	-.10	-.24
		Máximo	.25	-.30	.38	.34	.23	.34	.31	.23	.34

Nota: Os valores de alfa de Cronbach, se o item for eliminado, apresentados nos exames de Psi. Diferencial, 2010/11, dizem respeito às partes teóricas e prática separadamente e não ao total da escala

No Quadro 10, são apresentadas as correlações entre os itens de escolha múltipla, dicotómicos e não dicotómicos, e o item de desenvolvimento, com o intuito de depreender a relação entre os dois formatos de perguntas, através do coeficiente de

Pearson. Marôco (2011) considera que as correlações são fracas quando o valor absoluto é inferior a .25; moderadas entre .25 e .50; fortes entre .50 e .75; e muito fortes quando o valor é maior ou igual a .75.

Quadro 10 - Correlações entre Itens de Escolha Múltipla e Item de Desenvolvimento

	Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
Correlações	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Èp.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
T+P EM – T+P Des.	**.67**	**.64**	**.59**	-	-	-	-	-	-
EM – Des.	-	-	-	**.57**	.48	**.73**	**.68**	**.58**	**.50**
T+P EM dicotómico – T+P Des.	.45	**.53**	.41	-	-	-	-	-	-
EM dicotómico - Des.	-	-	-	.44	.39	**.61**	**.59**	.40	.44
Correlações entre os itens das Partes Teórica e Prática									
T EM – T Des.	**.56**	**.54**	**.55**	-	-	-	-	-	-
P EM – P Des.	.42	**.51**	.46	-	-	-	-	-	-
T EM – P Des.	.43	**.52**	.31	-	-	-	-	-	-
T Des. – P EM	**.52**	.44	.33	-	-	-	-	-	-
T EM – P EM	.28	.42	.28	-	-	-	-	-	-
T Des. – P Des.	**.68**	**.71**	**.60**	-	-	-	-	-	-

Valores a negrito representam correlações fortes ([.50, .75]) e muito fortes (≥.75), (Marôco, 2011)

As correlações entre itens de escolha múltipla, não dicotómicos, e desenvolvimento tomam valores entre .48 e .73, refletindo correlações moderadas a fortes, demonstrando que existe uma associação positiva entre as duas variáveis. Porém, as correlações entre itens de escolha múltipla dicotómicos e de desenvolvimento, apesar de moderadas representam correlações mais baixas, entre .39 e .61. O que sugere que os itens de escolha múltipla dicotómicos (quando não se considera a justificação) são menos consistentes com o item de desenvolvimento.

Também verificamos, no exame de Psi. Diferencial 2010/11, que as correlações entre itens de escolha múltipla e de desenvolvimento na mesma parte do exame são mais fortes (entre .42 e .56) do que entre as partes (entre .33 e .52). É ainda de assinalar que as correlações entre o item de desenvolvimento da parte Teórica (T Des.) e o item de desenvolvimento da parte Prática (P Des.) representam associações mais fortes (entre

.60 e .71), do que os itens de escolha múltipla da parte Teórica (T EM) com os da parte Prática (P EM), entre .28 e .42.

No Quadro 11, apresentam-se as correlações entre a pontuação obtida no conjunto dos itens dicotómicos, que equivale ao número total de respostas certas, e a pontuação obtida nos itens não dicotómicos, mas dela retirando a pontuação relativa à escolha das opções corretas (0.25 por cada item correto) – um procedimento necessário para que esta variável expresse apenas a qualidade das justificações das respostas corretas e não a sua correção, contemplada já na outra variável.

Quadro 11 - Correlações entre o número de itens corretos e a qualidade das justificações dadas aos itens corretos.

	Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
	1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
Correlações de Spearman entre as duas variáveis	**.72**	**.72**	**.89**	**.81**	**.84**	**.84**	**.80**	**.85**	**.89**

Nota: corrigido o caráter espúrio das duas variaveis, excluindo da segunda (qualidade das justificações) a pontuação da respsota correta (0.25) e retendo apenas a pontuação da justificação.
Valores a negrito representam correlações fortes ([.50, .75]) e muito fortes (≥.75), (Marôco, 2011)

As correlações obtidas variam de .72 a .89, demonstrando ser fortes a muito fortes (Marôco, 2011), o que constitui um potente indicador de consistência da avaliação de conhecimentos nas unidades curriculares sob análise: as duas variáveis, que refletem dois comportamentos distintos dos estudantes (o comportamento de seleção da alternativa correta e o de justificação desse seleção), estão positiva e muito significavamente correlacionadas, tendo elevado valor preditivo entre si. Embora tal resultado pudesse sugerir redudância da classificação das justificações (respostas breves) em relação à simples cotação dicotómica das respostas corretas, levando a possibilidade de dispensar essa tarefa de classificação, quando se passa ao nível da análise de itens, verificou-se que os itens de escolha múltipla com justificação apresentam um muito superior poder discriminativo sobre os dicotómicos (cf. Quadros 5 e 6). Assim, os resultados do Quadro 11 sublinham a consistência interna dos critérios de avaliação dos itens de escolha múltipla (precisão das medidas) enquanto o estudo da discriminação dos itens (Quadro 5 e 6) acentua a superior discriminação dos conhecimentos dos estudantes decorrente dos itens não dicotómincos (validade das

medidas), o que conduz a concluir favoravelmente em relação ao pedido de justifcação pelas opções de resposta, sendo altamente consistentes entre si.

No Quadro 12, encontram-se as correlações entre os dois tipos de itens e os respetivos totais, nas partes Teórica e Prática. Note-se que as correlações espúrias (contaminadas, devido a partilharem itens) foram corrigidas, mediante a aplicação da Fórmula de Correção de Correlações Espúrias de McNemar (McNemar, 1949 cit. por Marques, 1969, p.61).

Observando as correlações, nos exames de Psi. Diferencial, 2010/11, entre os itens de escolha múltipla e de desenvolvimento com o resultado total, verificamos que os itens de desenvolvimento, em ambas as partes (Teórica e Prática), apresentam correlações superiores (forte/muito forte), variando entre .68 e .82, quando comparadas com as correlações dos itens de escolha múltipla (com o total), que variam entre .54 e .67. Podemos inferir que os itens de desenvolvimento são bons preditores do resultado final, no que diz respeito aos exames de Psi. Diferencial, do ano letivo 2010/11, em todas as épocas de exame, ou pelo menos melhores preditores do que os itens de escolha múltipla, o que bem justifica que tal formato dos itens não seja abandonado.

Quadro 12 - Correlações entre partes Teórica e Prática e Totais

Correlações		Psicologia Diferencial (2010/11) (n=267)			Psicologia Diferencial (2011/12) (n=337)			Psicometria (2011/12) (n=321)		
		1ªÉpoca (n=93)	2ºÉpoca (n=118)	Ép.Esp. (n=56)	1ªÉpoca (n=179)	2ºÉpoca (n=112)	Ép.Esp. (n=46)	1ªÉpoca (n=151)	2ºÉpoca (n=116)	Ép.Esp. (n=54)
Teórica EM	T EM+Des.	**.56**	**.54**	**.55**	-	-	-	-	-	-
	P EM+Des.	.44	**.53**	.32	-	-	-	-	-	-
	T+P – EM	.28	.42	.28	-	-	-	-	-	-
	T+P – Des.	**.55**	**.57**	.49	-	-	-	-	-	-
	T+P	**.62**	**.67**	**.59**	-	-	-	-	-	-
Teórica Des.	T EM+Des.	**.56**	**.54**	**.55**	-	-	-	-	-	
	P EM+Des.	**.72**	**.64**	**.55**	-	-	-	-	-	-
	T+P – EM	**.68**	**.57**	**.57**	-	-	-	-	-	-
	T+P – Des.	**.68**	**.71**	**.60**	-	-	-	-	-	-
	T+P	**.82**	**.76**	**.75**	-	-	-	-	-	-
Teórica EM+Des.	P EM	.47	.49	.32	-	-	-	-	-	-
	P Des.	**.65**	**.71**	**.52**	-	-	-	-	-	-
	P EM+Des.	**.68**	**.67**	**.50**	-	-	-	-	-	-
	T+P – EM	**.84**	**.78**	**.78**	-	-	-	-	-	-
	T+P – Des.	**.86**	**.87**	**.80**	-	-	-	-	-	-
	T + P	**.68**	**.67**	**.50**	-	-	-	-	-	-
Prática EM	P EM+Des	.42	**.51**	.46	-	-	-	-	-	-
	T+P – EM	.28	.42	.28	-	-	-	-	-	-
	T+P – Des.	**.52**	**.51**	.44	-	-	-	-	-	-
	T+P	**.59**	**.64**	**.54**	-	-	-	-	-	-
Prática Des.	P EM+Des.	.42	**.51**	.46	-	-	-	-	-	-
	T+P – EM	**.53**	**.61**	.49	-	-	-	-	-	-
	T+P – Des.	**.68**	**.71**	**.60**	-	-	-	-	-	-
	T+P	**.75**	**.79**	**.68**	-	-	-	-	-	-
Prática EM+Des.	T+P – EM	**.74**	**.87**	**.72**	-	-	-	-	-	-
	T+P – Des.	**.88**	**.80**	**.78**	-	-	-	-	-	-
	T+P	**.68**	**.67**	**.50**	-	-	-	-	-	-
T+P – EM	T+P	**.67**	**.64**	**.59**	**.57**	.48	**.73**	**.68**	**.58**	**.50**
T+P – Des.	T+P	**.67**	**.64**	**.59**	**.57**	.48	**.73**	**.68**	**.58**	**.50**

Valores a negrito representam correlações fortes ([.50, .75]) e muito fortes (≥.75), (Marôco, 2011)

Estando realizadas todas as análises relacionadas com a qualidade dos itens, consistência interna e valor preditivo/validade dos itens e das partes constituintes dos exames, procurou-se ainda explorar os percursos dos estudantes repetentes – analisar a sua evolução e proceder à comparação dos resultados nos dois formatos de avaliação de aprendizagens, na unidade curricular: Psi. Diferencial.

O quadro referente a esta análise está apresentado em anexo (Anexo 2 – Quadro 16), onde são apresentadas as médias, desvios-padrão, mínimos e máximos dos examinados repetentes nos exames de Psi. Diferencial nos dois anos letivos (2010/11 e 2011/12), bem como o número de exames que realizaram, a média no primeiro formato avaliativo (Psi. Diferencial, 2010/11, nas três épocas) e no segundo formato (Psi. Diferencial, 2011/12, nas três épocas).

Na amostra de alunos repetentes (167 alunos), dos exames de Psi. Diferencial (em 2010/11 e 2011/12), 32 realizaram pelo menos um exame, desta unidade curricular, em cada ano letivo, e estes foram os retidos nesta análise. No quadro, observamos que a média dos resultados totais nos exames de Psi. Diferencial, varia entre 2,06 valores e 8,42 valores (na escala de 0 – 20). Para além disso, constatamos que a média do primeiro formato de exame (Psi. Diferencial, 2010/11) é geralmente menor que a média do segundo formato de exame, refletindo que, para os alunos repetentes, o segundo formato foi mais favorável. Desta forma, utilizámos um teste não paramétrico para amostras emparelhadas (Teste de Wilcoxon – em anexo 2, Quadro 17), com o intuito de verificar se existe uma diferença significativa entre os resultados dos dois formatos. Encontrou-se uma diferença significativa a favor do segundo formato, uma vez que se verificou um maior número de observações com médias superiores nos exames de Psi. Diferencial, de 2011/12, do que nos exames de Psi. Diferencial, de 2010/11.

Para analisar os resultados totais nos diferentes exames de Psi. Diferencial (três épocas, e dois anos letivos, 2010/11 e 2011/12) de forma mais fina, procedeu-se a uma análise qualitativa ao nível intraindividual, com o intuito de compreender a tendência geral na variação das classificações dos examinados que reprovaram, ao longo do tempo. Percebemos que 20 dos 167 estudantes que reprovaram alguma vez nos exames de Psi. Diferencial (em ambos os anos letivos), repetiram os exames 4 a 5 vezes (11 e 9 estudantes, respetivamente). Nesta amostra, verificamos que os estudantes que realizaram 4 exames de Psi. Diferencial obtêm classificações que variam com maior frequência entre 3 e 8 valores, tendo uma média de 5,70; nos estudantes que realizaram

5 exames de Psi. Diferencial, as classificações variam com maior frequência entre 4 e 7 valores, tendo uma média de 6,24. Nos examinados que realizaram três exames de Psi. Diferencial, verifica-se uma média de 6,76; enquanto, os que realizaram apenas dois exames, a média dos resultados totais é a mais alta, de 8,17 valores. Podemos inferir que quando os examinados repetem mais do que dois exames, têm uma maior tendência em obter classificações mais baixas, não melhorando o seu desempenho sensivelmente ao longo do tempo, e consequentemente uma maior dificuldade em alcançar classificações positivas na unidade curricular, o que mostra algum sentido lógico, uma vez que ambos os indicadores, número de repetições do exame e nível de desempenho (classificação), se mostram coerentes (cf. Anexo 2 – Quadro 19). Verifica-se também que o exame de Psi. Diferencial, do ano letivo 2011/12, de 1ªépoca, foi mais difícil para os examinados repetentes, uma vez que nenhum consegue obter uma classificação positiva, neste exame.

Esta análise foi realizada também para os examinados que realizaram mais do que um exame de Psicometria, no mesmo ano letivo, 2011/12, em diferentes épocas. O quadro é apresentada em anexo (Anexo 2 – Quadro 18) e verifica-se que nesta amostra, constituída por 52 estudantes que realizaram os exames 2 ou 3 vezes, a média das classificações foi de 7,36 valores e as classificações com maior frequência se situaram entre 6 e 8 valores.

V. DISCUSSÃO DE RESULTADOS

Ao analisarmos dois formatos de avaliação de conhecimentos para a mesma unidade curricular (Psi. Diferencial), deparamo-nos com algumas diferenças, dignas de atenção, nomeadamente no que diz respeito às médias totais obtidas, que refletem o desempenho dos examinados na unidade curricular: verificou-se genericamente um resultado favorável ao segundo formato de exame. Apesar de o número de participantes, em cada época de exame, ser variável, o primeiro formato, que dividia o exame em duas partes, Teórica e Prática, avaliando, assim, conhecimento teóricos e práticos em segmentos distintos da prova, é tendencialmente mais difícil para os examinados. Para além disso, o segundo formato de exame introduz uma maior variabilidade nos dados (maior desvio-padrão), o que nos leva a concluir que também deverá diferenciar melhor os examinados uns dos outros.

Este facto leva-nos a ponderar que o primeiro formato de exame poderá tornar o planeamento e organização do estudo mais complexo. Ou seja, ao "obrigar" os examinados a separar conteúdos da mesma unidade curricular, poderá induzir os estudantes a focarem-se muito numa parte (Teórica ou Prática) em detrimento de outra, consequentemente o desempenho no exame tenderá a ser lesado e os estudantes poderão estar mais informados sobre uma parte dos conteúdos. No segundo formato, o planeamento do estudo envolve a interligação de conteúdos, uma vez que o exame avalia também conhecimentos teórico-práticos e, desta forma, os examinados alcançam talvez maior *insight* e juízo critico, o que poderá resultar em desempenhos mais favoráveis.

Particularmente notáveis foram, os resultados quanto à qualidade dos itens de escolha múltipla, nos exames de ambas as unidades curriculares sob análise, os quais foram genericamente positivos, se devidamente enquadrados na literatura.

Scialfa et al. (2001) apontam para um valor ótimo de índice de dificuldade de .50, porém Sax (1980) e Colbert (2011) mencionam que apesar de este ser o valor ideal, não tem em conta os efeitos do "*guessing*" na resposta ao item, assinalando um valor de .67 como o mais indicado. Por seu lado, Kline (2005) aponta para um intervalo entre .20 e .80 como adequado para o índice de dificuldade. Na análise deste índice nos 9 exames sob análise, encontrámos valores médios do índice de dificuldade entre .48 e .65, refletindo uma dificuldade adequada para a avaliação de conhecimentos nas unidades

curriculares estudadas (sendo que 88% dos itens de escolha múltipla se encontram dentro do intervalo de .20 a .80, assinalado por Kline, 2005).

Os coeficientes médios de discriminação, quando se considera os itens de escolha múltipla com justificação (itens não dicotómicos), à exceção dos exames de 1ª época e época especial de Psi. Diferencial de 2010/11, expressam um bom poder discriminativo, apresentando resultados de .29 e .39. DiBatista e Kurzawa (2011) referem que a literatura sugere, para exames realizados em sala de aula, um intervalo satisfatório entre .20 e .30., o que é concordante com o seu estudo, onde analisaram 1198 itens de escolha múltipla, e o coeficiente médio de discriminação encontrado foi de .25. Isto leva-nos a ponderar que, mesmo com alguns itens de escolha múltipla com índices de discriminação abaixo do aceitável (<.20), os exames construídos para avaliação das aprendizagens em ambas as unidades curriculares, são genericamente discriminativos das diferenças individuais entre examinados, quanto aos conhecimentos sob avaliação. Para além disso, de um modo geral, verificamos, de novo, uma diferença entre os dois formatos de avaliação de conhecimentos, uma vez que no primeiro formato dois dos três exames apresentam coeficientes médios de discriminação pouco satisfatórios, e evidenciam vários itens de escolha múltipla com índices de discriminação negativos e/ou próximos de 0. Demonstra-se, assim, um contraste entre os dois formatos considerados, favorável ao segundo, visto este último, em ambas as unidades curriculares, apresentar consistentemente índices de discriminação aceitáveis ou mesmo bons (cerca de 72% dos índices no segundo formato de exame tem valores dicriminativos aceitáveis/bons).

Outro elemento importante a ser discutido é a qualidade dos distratores, onde os resultados encontrados, no presente trabalho, apontam para cerca de 89% de distratores eficientes. DiBattista e Kurzawa (2011) analisaram 3819 distratores, constituintes de itens de escolha múltipla de 16 exames de conhecimentos, e apuraram apenas cerca de 55% dos mesmos como eficientes (respeitando os dois critérios anteriormente descritos: o distrator deverá ser selecionado por mais de 5% dos examinados e o grupo com melhores classificações, nos itens de escolha múltipla, deverá selecionar menos vezes o distrator do que o grupo com piores classificações). Outros estudos referidos pelos mesmos autores constatam que a maioria dos distratores que são utilizados em avaliações das aprendizagens funcionam de forma muito pobre, o que coloca em evidência a dificuldade por parte dos professores em construir distratores eficazes

(respostas erradas plausíveis mas inequívoca e objetivamente erradas). Assim sendo, podemos concluir, com clareza, que a maioria dos distratores construídos para a avaliação de conhecimentos em ambas as unidades curriculares se mostrou muito apropriada, revelando que foram redigidos com cuidado e ponderação, seguindo as normas estabelecidas para a construção desse tipo de itens (Haladyna, 2004).

O estudo da consistência interna fornece ainda dados extremamente importantes sobre a qualidade dos exames. Ebel e Frisbie (1986) referem que a maioria dos exames realizados em sala de aula apresenta alfas de Cronbach menores ou iguais a .50, no entanto, são passiveis de alcançar valores mais altos. DiBattista e Kurzawa (2011), por seu lado, mencionam que para considerar um exame com uma consistência interna aceitável, deverá apresentar um valor mínimo de alfa de Cronbach de .70. Nas análises que realizámos, verifica-se que quatro dos nove exames atingem uma fiabilidade aceitável (≥.70), apresentando também correlações inter-item ≥.20. Constatamos que estes exames pertencem todos ao segundo formato, levando a considerar que este formato de avaliação de conhecimentos fornece uma maior consistência entre os itens. É de sublinhar que estes valores de alfa de Cronbach pertencem a uma escala constituída por apenas 10 itens de escolha múltipla, e é sabido que o baixo número de itens prejudica a consistência interna. À semelhança do que é sabido quanto à medição em psicologia, Ebel e Frisbie (1986) acrescentam mesmo que a fiabilidade de uma escala, em testes de conhecimentos, é maior quanto maior for o número de itens e, de facto, na maioria das publicações e investigações que consultámos, a análise dos itens de escolha múltipla é realizada com mais de 50 itens (DiBattista & Kurzawa; Ebel & Frisbie, 1986; Bacon, 2003). Para além disso, há que assinalar que os testes de conhecimentos têm forçosamente de abranger variedade nos conteúdos, de forma a aumentar a representatividade ou cobertura do domínio de conteúdo (a validade de conteúdo), o que, consequentemente, inviabiliza a obtenção de coeficientes alfa de Cronbach muito elevados, como os que se obtêm nas medidas de traços psicológicos, que incluem itens muito mais homogéneos no tipo de conteúdos, por vezes até sentidos pelos respondentes como muito repetidos ou redundantes (por exemplo: em escala de testes de personalidade, como uma escala de depressão).

Uma das análises mais importantes realizadas nesta monografia, passou pela comparação entre o funcionamento metrológico de itens de escolha múltipla com justificação (não dicotómicos) e sem justificação (itens dicotómicos). Foram poucos os

estudos que encontrámos que procedem a este tipo de comparação. Alguns autores apontam para o uso de escolha múltipla com justificação como optativo para os estudantes, isto é, os alunos utilizam essa opção quando percecionam o item como ambíguo ou confuso. Assim, o uso da escolha múltipla neste formato contribui para diminuir o *stress* e a ansiedade nos alunos, e cria um maior diálogo entre estudantes e professores (Dood & Leal, 1998; Halaydna, 2004). No entanto, Tamir (1991) defende que o uso de justificação nos itens de escolha múltipla é extremamente importante, uma vez que proporciona a oportunidade de os examinados explicarem a sua escolha, expressando a sua posição acerca do conteúdo, fornecendo informação ao professor sobre as conceções dos estudantes. Este autor realizou um estudo comparando as percentagens de estudantes que escolhiam a alternativa correta nos itens de escolha múltipla e os que forneciam uma justificação satisfatória sobre o mesmo item, e encontrou diferenças significativas, levando-o a concluir que um número considerável de estudantes que escolhiam a alternativa correta não compreendiam totalmente o conteúdo do item (Tamir, 1991).

As análises levadas a cabo nesta investigação, evidenciaram que existe uma diferença considerável na consistência interna do exame e no seu poder de discriminação do nível de conhecimentos dos estudantes, quando, nos itens de escolha múltipla, se toma apenas em consideração o acerto ou erro na seleção da alternativa de resposta (itens dicotómicos) ou quando se toma em conta, também, a justificação da opção (itens não dicotómicos). Verificou-se que a consistência interna é consideravelmente inferior, nos itens dicotómicos, dado que os valores de alfa de Cronbach descem em média .25, e as correlações inter-itens também são menores. Para além disso, o coeficiente médio de discriminação apresenta resultados inferiores, entre -.06 e .23, e constata-se a existência de vários itens com muito fraco poder discriminativo. Quando observamos as correlações entre os itens de escolha múltipla dicotómicos e não dicotómicos e o item de desenvolvimento, constatamos, ainda, que são mais fortes entre os itens não dicotómicos e o item de desenvolvimento, o que constitui novo indício favorável aos itens não dicotómicos. Acresce que as correlações elevadas, nos itens de escolha múltipla, entre o número total de respostas corretas e a qualidade da justificação das opções de resposta constituem um indicador da consistência dos critérios que presidem à avaliação de conhecimentos, sendo que a análise de itens acentou ainda a superioridade dos itens com justificação, do ponto de vista da validade das medidas que

proporcionam (poder discriminativo). Este conjunto de resultados leva-nos a concluir que para a avaliação de conhecimentos, nas duas unidades curriculares sob análise, o uso de itens de escolha múltipla com justificação favorece não só a fiabilidade do exame (maior consistência interna), mas também discrimina de forma mais eficiente o nível de conhecimento e o domínio dos conteúdos de Psi. Diferencial e Psicometria, entre os examinados (superior validade).

Quanto à análise realizada com a amostra de examinados repetentes, na mesma unidade curricular (Psi.Diferencial), os resultados apontam, de novo, para o segundo formato de exame como mais acessível, nos exames de Psi. Diferencial, dado as médias serem consistentemente mais altas. Verifica-se que os examinados que repetem mais do que duas vezes o exame de Psi. Diferencial, têm maior dificuldade em alcançar sucesso na disciplina, o que coloca em relevo a coerência entre o número de repetências e o nível médio das classificações. Possivelmente, a frustração em não conseguir obter sucesso, que poderá conduzir à desmotivação, resultando num estudo menos sistemático, planeado, e/ou investido sejam fatores que favorecem estes resultados. Ayodele e Adebiyi (2013) afirmam que quando um estudante obtém insucesso de forma consistente em avaliações sucessivas, irá desenvolver uma baixa autoestima e falta de confiança para estudar e, consequentemente, terá maior dificuldade em obter sucesso posteriormente. Embora estes resultados sublinhem a necessidade de tomar medidas de intervenção remediativas com estes estudantes, não deixam de evidenciar a qualidade metrológica e poder discriminativo destes exames, enquanto instrumentos de avaliação de conhecimentos, ao revelarem o seu poder de identificar, de forma sistemática e consistente, os mesmos estudantes e ao apresentarem um padrão lógico de resultados, com coerência entre o número de repetições e o nível das classificações. Constatamos que a correlação entre o número de repetições e o nível das classificações nos exames de Psi. Diferencial, de 2010/11 é -.13, de Psi.Diferencial, de 2011/12 é -.40 e de Psi. Diferencial, de 2010/11 + 2011/12, é -.37 (cf. Anexo 2 – Quadro19), estes dados são indicadores de consistência nas classificações, uma vez que os alunos que consistentemente obtêm piores resultados são os que mais vezes realizam o exame. Contudo, é de salientar que a menor coerência verificada entre as duas variáveis, encontra-se no exame de Psi. Diferencial, 2010/11, o que possivelmente se deverá às qualidades psicométricas assinaladas neste trabalho, bem como devido ao sucesso no

exame, neste ano letivo, ser condicionado pela obtenção de uma classificação positiva em ambas as partes (≥9.5 valores).

Posto isto, as análises levadas a cabo neste trabalho demonstram que os itens de escolha múltipla construídos com cuidado e ponderação, no quadro das normas orientadoras existentes para esse efeito, permitem obter bons resultados nos três critérios principais propostos pela maioria dos autores consultados (Ebel & Frisbie, 1986; Halaydna, 2004; Sax, 1980; DiBattista e Kurzawa, 201; Bacon, 2003; Kline, 2005; Scialfa et al., 2001): índices de dificuldade ≥.50, índice médio de discriminação entre . 29 e .39, e 89% dos distratores construídos de forma eficaz. A consistência interna apesar de baixa em alguns dos exames (≤.50), atinge valores aceitáveis (≥.70), em 4 dos 6 exames, com o segundo formato, refletindo a possibilidade de aumentar a fiabilidade da escala, utilizando o segundo formato.

Podemos observar, nos formatos de avaliação de conhecimentos de Psi. Diferencial, no ano letivo 2010/11, que os itens de desenvolvimento apresentam um ótimo poder preditivo, uma vez que obtém correlações com os resultados totais entre .68 e .82. Também obtemos correlações moderadas a fortes entre os itens de escolha múltipla e de desenvolvimento, em todos os exames, variando entre .48 e .73, o que demonstra que os dois tipos de perguntas estão correlacionadas de forma positiva, como seria desejável, avaliando assim ambas o mesmo domínio de conhecimentos, mas não apresentando correlações de tal modo elevadas, que sugerissem ser redundantes, dispensando-se aplicação, de dois formatos de itens. Pelo contrário, as correlações positivas, mas moderadas sugerem a utilidade de manter a diversidade de formatos de avaliação, o que poderá ser também favorável a uma maior quantidade de estudantes, não facilitando a tarefa exclusivamente aos que melhor dominem apenas um dos formatos de resposta (só desenvolvimento ou só escolha múltipla).

Atendendo aos objetivos inicialmente propostos e aos resultados das análises efetuadas ao longo deste trabalho, concluímos que o segundo formato de exame para avaliação das aprendizagens das unidades curriculares de Psi. Diferencial e de Psicometria diferencia melhor os examinados e avalia de forma mais coerente os conteúdos pretendidos. Para além disso, foram encontradas fortes evidências (correlações fortes a muito fortes entre qualidade da justificação e itens de escolha múltipla dicotómicos e maior poder discriminativo dos itens de escolha múltipla não dicotómicos) de que o uso de itens de escolha múltipla com justificação diferencia de

forma eficiente os conhecimentos dos examinados, proporcionando uma melhor qualidade da avaliação, uma vez que quando o examinado justifica a escolha da alternativa põe em evidência os conhecimentos e o seu juízo e pensamento crítico.

VI. CONCLUSÃO

O presente trabalho sublinha a utilidade e a necessidade de recorrer a estudos docimológicos, que coloquem em evidência as potencialidades e limitações dos exames de avaliação de conhecimentos, com o intuito de melhorar a qualidade dos métodos de avaliação. Os resultados do presente estudo, em particular, serão levados em conta na construção de posteriores instrumentos de avaliação de conhecimentos, em ambas as unidades curriculares estudadas. Verificamos de forma sistemática que os instrumentos de avaliação de aprendizagens utilizados em sala de aula, particularmente nas universidades, são utilizados de forma a determinar o sucesso/insucesso escolar, o que devia tornar imprescindível o seu estudo científico. Neste sentido, mostra-se útil revisitar os conceitos e as metodologias utilizadas pela docimologia, e aplicá-los aos instrumentos de avaliação de aprendizagens, ainda que estes possam ser hoje usados em contexto e com finalidades consideravelmente distintas, muito para lá da mera avaliação sumativa. No presente estudo, constatou-se que os instrumentos que estão subjacentes a um ensino centrado na função formativa, e ultrapassando o carácter meramente instrumental que se encontrava nas investigações docimológicas, de outrora, são passíveis de ser submetidos, mais, deveriam ser submetidos a escrutínio científico, com o intuito de monitorizar o seu valor, enquanto instrumentos de avaliação das aprendizagens, contornando como tal algumas das críticas feitas à docimologia clássica (Correia, 2002; Despresbiteris, 1998; Leclercq, Nicaise & Demeuse, 2004). Por outras palavras, a utilização formativa, e não apenas a sumativa, destes instrumentos não pode, de modo algum, dispensar a averiguação da sua qualidade metrológica e do seu valor à luz dos critérios científicos.

Em conclusão, tomando em consideração os objetivos propostos inicialmente no presente estudo exploratório podemos concluir que o segundo formato de avaliação de conhecimentos para as unidades curriculares de Psi. Diferencial e Psicometria, é adequado, visto apresentar qualidades psicométricas bem enquadradas nos critérios definidos na bibliografia consultada (DiBatista & Kurzawa, 2011; Ebel & Frisbie, 1986; Bacon, 2003; Kline, 2005; Haladyna, 2004; Sax, 1980). É de sublinhar que os três critérios para a análise de itens de escolha múltipla explorados (DiBatista & Kurzawa, 2011; Ebel & Frisbie, 1986; Haladyna, 2004), revelaram resultados com bastante qualidade, possivelmente decorrentes de uma construção cuidada dos itens, orientada por exigentes princípios técnicos.

Embora, a consistência interna, no segundo formato de avaliação de conhecimentos, pudesse ser aumentada, acrescentando mais itens de escolha múltipla, uma tal opção implicaria prolongar o tempo de exame, que é já de 2horas e 30 minutos, sendo que, ainda assim, a necessidade de cobertura de uma diversidade de conteúdos, neste tipo de exames, dificultaria, mesmo com mais itens, a obtenção de elevados coeficientes de alfa de Cronbach. Por outro lado, é de sublinhar que os dois tipos de questões, itens de escolha múltipla com justificação e de desenvolvimento, obtêm correlações moderadas entre si, sugerindo a utilidade de manter a diversidade de formatos dos itens na avaliação de conhecimentos, ao invés de construir testes exclusivamente com itens de escolha múltipla.

Ao se verificarem, em todos os exames, correlações elevadas entre a qualidade da justificação dada aos itens de escolha múltipla e o número de respostas acertadas, pode concluir-se que as justificações das opções de resposta aos itens conferem maior robustez (precisão) a toda a avaliação, ao fornecer um indicador alternativo consistente, que se espera convergente com o número de itens acertados, o qual permite confirmar, em cada caso, o nível de conhecimentos evidenciado pelo estudante. Por outro lado, o evidente superior poder discriminativo dos itens considerando a justificação demonstrou que este formato de resposta proporciona uma medida mais válida, do que a proporcionada pelos itens dicotómicos, dos conhecimentos avaliados nas unidades curriculares sob análise.

Por fim, a análise realizada com a amostra dos estudantes repetentes, evidenciou que um maior número de repetições está associado a uma média mais baixa de classificações. Apesar de estes resultados que apresentarem coerência lógica, e indicarem alguma consistência nos critérios de classificação dos estudantes ao longo dos exames, não dispensam que se tomem medidas para identificar o tipo de dificuldades enfrentadas por estes estudante e tomar medidas como: encorajá-los a assistir de forma assídua às aulas, a esclarecer dúvidas e obter orientações para estudo nos tempos de apoio tutorial e a receber *feedback*, junto das docentes, de forma a compreender os seus pontos fortes e fracos, com o intuito de aumentar a sua motivação e investimento nas unidades curriculares.

Apesar das conclusões a que permitiu chegar, algumas limitações, há a apontar a este estudo, primeiro, não se ter alcançado todos os objetivos ambicionados num primeiro momento, que iriam tornar a análise mais rica. De facto, o levantamento,

previsto no início, das médias de ingresso no Mestrado Integrado em Psicologia (MIP) (notas de candidatura à Universidade), bem como das médias de acesso ao 2ºciclo de MIP e das classificações dos estudantes noutras unidades curriculares obrigatórias, iria enriquecer esta análise, sobretudo do ponto de vista da averiguação da validade das avaliações proporcionadas por estes exames na predição de critérios externos. Porém, não foi possível explorar estes dados, por um lado, dada a já extensa e exigente análise realizada neste trabalho e, por outro lado, porque, com o intuito de garantir o anonimato dos participantes, não foi possível a investigadora ter acesso direto aos dados, dependendo esse acesso a devida transformação dos números de estudantes em números convencionais de participante, tarefa que não foi possível à docente realizar em tempo útil. Assim, propõe-se a continuação desta investigação, acrescentando a exploração das relações entre os vários resultados dos exames e uma diversidade de critérios externos, ou mesmo a recolha de outras evidências de validade.

Uma segunda limitação que poderá ser identificada remete para o modelo de medida clássico (também conhecido como "modelo do resultado verdadeiro") subjacente a toda a análise metrológica. Seria útil ensaiar, em alternativa ou mesmo em complementaridade, a aplicação de um modelo de traço latente (Teoria de Resposta ao Item ou TRI), uma metodologia que oferece algumas importantes vantagens sobre os métodos clássicos de análise, mas que requer o uso de programas informáticos não disponíveis para os estudantes e de aquisição dispendiosa. Deixa-se, contudo, assinalada a intenção de futuramente recorrer a tal metodologia, o que abrirá até a possibilidade de vir a construir novos exames a partir de uma *pool* de itens selecionados a partir de estudo prévio e, como tal, de propriedade metrológicas conhecidas.

No seu conjunto, este trabalho permitiu concluir muito favoravelmente, em relação à pertinência e aplicabilidade da metodolodia proposta para o estudo sistemático da qualidade dos exames escritos, o que sugere que seria útil ser ensaiada noutras unidades curriculares, de modo a verificar e eventualmente promover a qualidade dos instrumentos de avaliação das aprendizagens em uso. Uma tal ótica de investigação mostra-se, de facto, de maior relevância, já que os exames escritos, enquanto insturmentos de avaliação de conhecimentos mais disseminados no ensino superior, são absolutamente cruciais no processo de ensino-aprendizagem, na determinação do sucesso ou do insucesso escolar e na definição do futuro percurso académico e vocacional dos estudantes.

VII. REFERÊNCIAS

Afonso, A. F. (2011). *Concepções e práticas de avaliação de professores de Ciências da Natureza do 2º Ciclo do Ensino Básico: Um olhar dirigido para os testes de avaliação*. Dissertação de Mestrado, Instituto Politécnico de Bragança – Escola Superior de Educação de Bragança, Portugal.

Albuquerque, T. S. & Oliveira, E. S. (2012). *Avaliação da Educação e da aprendizagem*. Curitiba: IESDE Brasil.

Ayodele, C.S. e Adebiyi, D.R. (2013). Study habits as influence of academic performance of university undergraduates in Nigeria. *Research Journal in Organizational Psychology & Educational Studies, 2* (3), 72-7. Retirado de: http://rjopes.emergingresource.org/articles/STUDY%20HABITS%20NEW.pdf

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 2* (1), 31-36. DOI: 10.1177/0273475302250570

Benavente, A. (1990). Insucesso escolar no contexto português – abordagens, concepções e políticas. *Análise Social, vol. XXV* (108-109), 71-733.

Bernheim, C. T. & Chauí, M. S. (2008). *Desafios da universidade na sociedade do conhecimento: Cinco anos depois da conferência mundial sobre a educação superior*. Brasília: UNESCO. Retirado de: http://unesdoc.unesco.org/images/0013/001344/134422por.pdf

Bisinoto, C. Marinho, C. & Almeida, L. (2010). Contribuições da Psicologia Escolar à promoção do sucesso académico na educação superior. In *Seminário Internacional "Contributos da Psicologia em Contextos Educativos*", I (p. 102-116). Braga: Universidade do Minho

Bittencourt, H. R., Creutzberg, M., Rodrigues, A. C., Casartelli, A. O. & Freitas, A. L. (2011). Desenvolvimento e validação de um instrumento para avaliação de disciplinas na educação superior. *Estudos em Avaliação Educacional, 22* (48), 91-114

Chabot, J. M. (2004). Evaluer, corriger, pondérer, noter, classer. *La Reveu du praticien, 54*, 311-312.

Cerny, R. Z. & Ern, E. (2001). *Uma reflexão sobre a avaliação formativa na educação à distância.* 24ª Reunião anual da Associação Nacional de Pós-Graduação e Pesquisa em Educação, Caxambu. Retirado de: http://www.cridi.ufba.br/twiki/pub/GEC/TrabalhoAno2001/uma_reflexao_sobre_a_avaliacao_formativa_na_ead.pdf

Colbert, M. A. (2001). *Statistical analysis of multiple-choice testing.* Alabama: Air command and staff college air university Maxwell AFB.

Correia, E. S. L. (2002). *Avaliação: Gerações da Avaliação – Traços Históricos*. Portugal: Universidade de Aveiro.

Cortesão, L. (2005). Formas de ensinar, formas de avaliar. Breve análise de práticas correntes de avaliação. *In Reorganização Curricular do Ensino Básico – Avaliação das aprendizagens: das concepções às práticas* (p. 37-42). Lisboa: Ministério da Educação.

Costa, M. G. (2007). *A avaliação nas séries iniciais do ensino fundamental*. Dissertação de Mestrado em Psicopedagogia. Universidade Cândido Mendes, Rio de Janeiro.

Curado, A. P. e Machado, J. (2005). *Percursos escolares dos estudantes da Universidade de Lisboa: Factores de sucesso e insucesso escolar na Universidade de Lisboa*. Lisboa, Universidade de Lisboa. Retirado de: http://repositorio.ul.pt/bitstream/10451/2996/1/9729086117.pdf

Decreto-Lei nº139/2012, de 5 de julho. *Diário da República, 1.ª série — N.º 129.* Ministério da Educação e da Ciência

Despacho normativo nº24-A/2012, de 6 de dezembro. *Diário da República, 2.ª série — N.º 236*. Ministério da Educação e da Ciência

DiBattista, D. & Kurzawa, L. (2001). Examination of the quality of multiple-choice items on classroom test. *The Canadian Journal of the Scholarship of Teaching and Learning, 2* (4). DOI: 10. 5206/cjsotl-rcacea.2011.2.4. Retirado de: http://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1061&context=cjsotl_rcacea

De Landsheere, G. (1976). *Avaliação contínua e exames: noções de docimologia*. Coimbra: Livraria Almedina.

Despresbiteris, L. (1998). Avaliação da aprendizagem do ponto de vista técnico-científico e filosófico-político. *Série Ideias, 8*, 161-172. São Paulo.

Dias, E. G. (2011). *Avaliação e (in)sucesso escolar. Estudo Exploratório*. Dissertação de Mestrado em Ciências da Educação. Instituto de Educação, Universidade do Minho.

Dood, D. K. & Leal, L. (1998). Answer justification: Removing the "Trick" from multiple-choice questions. *Teaching Psychology, 15* (1), 37-38. DOI: 10.1207/s15328023top1501_8.

Ebel, R. L. & Frisbie, D. A. (1986). *Essentials of educational measurement*. USA: Prentice-Hall

Estima, H. M. (2011). *O exame de matemática e as práticas de ensino e avaliação no 12ºano: perspetiva dos alunos*. Dissertação de Mestrado em Ciências da Educação. Instituto de Educação, Universidade de Lisboa.

Fernandes, D. (2004). *Avaliação das aprendizagens: Uma agenda, muitos desafios*. Cacém: Texto Editores.

Fernandes, D. (2006). Vinte anos de investigação das aprendizagens: Uma síntese interpretativa de artigos publicados em Portugal. *Revista Portuguesa de Pedagogia*, *40* (3), 289-348.

Fernandes, D. (2011). Articulação Da Aprendizagem, Da Avaliação E Do Ensino: Questões Teóricas, Práticas e Metodológicas. In *Do currículo à avaliação, da avaliação ao currículo*, ed J. M. Deketele e M. P. Alves, 131-142- Porto: Porto Editora.

Freitas, L. C., Sordi, M. R. L., Malavasi, M. M. & Freitas, H. C. L. (2009). Avaliação educacional: caminhando pela contramão. Petrópolis, RJ: Vozes.

Garcia, J. (2009). Avaliação e aprendizagem na educação superior. *Estudos sobre a Avaliação Educacional, 20* (43), 201-213.

Guba, E. G. & Lincoln, I. S. (1989). *Fourth generation evaluation*. Newbury Park, California: Sage Publications.

Hadji, C. (1994). *A avaliação, regras do jogo: das intenções aos instrumentos*. Porto: Porto Editora

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. USA: Lawrence Erlbaum Associates

Hoy, A. W. (2002). Educational Psychology. In *Encyclopedia of Education* (2nd Ed), J.W. Guthrie. (p. 7-683). London: Macmillan.

Kline, T. J. B. (2005). *Psychological testing: A pratical approach to design and evaluation*. Thousand Oaks: Sage

Leclercq, D., Nicaise, J. e Demeuse, M. (2004). Docimologie critique: des difficultés de noter des copies et d'attribuer des notes aux élèves. In M. Demeuse (ed), *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation*. Liège: Éditions de l'Université de Liège.

Lee, H., Liu, O. L. e Linn , M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*, 115-136. DOI: 10.1080/08957347.2011.554604

Luckesi, C. C. (2002). Avaliação da aprendizagem na escola e a questão das representações sociais. *Revista Científica, 4* (2), 79-88.

Marques, J. F. (1969). *Estudos sobre a Escala de Inteligência de Wechsler para Crianças (WISC). Sua adaptação e aferição para Portugal*. Lisboa: Instituto de Alta Cultura.

Marôco, J. (2011). *Análise estatística com o SPSS Statistics* (5ed). Lisboa: ReportNumber.

Marôco, J. & Garcia-Marques, T. (2006). Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas?. *Laboratório de Psicologia, 4* (1), 6-90.

Miranda, M. J. (1982). A Docimologia em perspetiva. *Revista da Faculdade da Educação, 8* (1), 39-69.

Simkin, M. G. & Kuechler, W. L. (2005). Multiple-choice test and student understanding: What is the connection?. *Decision Sciences Journal of Innovative Education, 3 (*1), 73-97.

Stuffebeam, D. L., Madaus, G. F. & Kellaghan, T. (2000). *Evaluation models: Viewpoints on educational and human services evaluation* (2ed). Boston: Kluwer-Nijhoff Publishing.

Martins. C. M. (2008). *Dez anos de investigações em avaliação das aprendizagens: Reflexões a partir da análise de dissertações de mestrado*. Dissertação de Mestrado em Ciências da Educação. Faculdade de Psicologia e Ciências da Educação, Universidade de Lisboa.

Nevo, D., Alkin, M. & Cartstensen, C. (1975). *Studies in educational evaluation*. Amsterdam: Elsevier Science.

Noizet, G. & Caverni, J. (1985). *Psicologia da avaliação escolar*. Coimbra: Coimbra Editora.

Pacheco, J. (1995). A avaliação dos alunos: algumas reflexões com os professores. In *Actas do Seminário Avaliação dos alunos dos ensinos básicos e secundário* (p. 7-14). Guimarães: Centro de Formação de Professores Francisco de Holanda Retirado de

Piéron, H. (1974). *Ciência e técnica dos exames* (2ed). Lisboa: Moraes.

Popham, W. J. (1975). *Educational evaluation*. New Jersey: Prentice-Hall.

Ramraje, S.N. & Sable, P. L. (2011). Comparison of the effect of post-instruction multiple-choice and short-answer test on delayed retention learning, *Australasian Medical Journal*, 4 (6), 332-339. DOI: 10.4066/AMJ.2011.727

Rehem, C. C. & Melo, M. A. (2008). Avaliação da aprendizagem no ensino superior: novos discursos e velhas práticas. *Revista de Educação PUC-Campinas, 25*, 59-65.

Ribeiro, L. C. (1991). *Avaliação da Aprendizagem* (3ed). Lisboa: Texto Editora.

Rosales, C. (1992). *Avaliar é refletir sobre o ensino*. Lisboa: Edições ASA

Santos, L.P. (2012). Implicações das práticas avaliativas no ensino superior na formação docente. *Revista de Educação, Linguagem e Literatura da UEG, 4* (2), 69-88.

Santos, M. R. e Varela, S. (2007). A avaliação como um instrumento diagnóstico da construção do conhecimento das séries iniciais do ensino fundamental. *Revista Eletrónica de Educação, 1* (1), 1-14.

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation* (2ed). Belmont: Wadsworth.

Scialfa, C., Legare, C., Wenger, L. & Dingley, L. (2001). Difficulty and Discriminability of introductory psychological test items. *Teaching of Psychology, 28* (1), 11-15.

Sobrinho, J. D. (2010). Democratização, qualidade e crise da educação superior: faces da exclusão e limites da inclusão. *Educação & Sociedade, 31* (113), 1223-1245

SPSS (2011). SPSS Statistics for Windonws, Version 20.0. Armonk N.Y.: IBM Corp.

Struyven, K. Dochy, F. & Janssens, S. (2005). Students' perception about evaluation and assessment in higher education: a review. *Assessment & Evaluation in Higher Education, 30* (4), 331-347. DOI: 10.1080/0260293042000318091

Tamir, P. (1991). Multiple-choice items: How to gain the most out of them. *Biochemical Education, 19* (4), 188-192.

Valadares, J. & Graça, M. (1998). *Avaliando para melhorar a aprendizagem*. Lisboa: Plátano.

Vianna, H. M.(1998). Avaliação educacional: vivência e reflexão. *Estudos sobre Avaliação Educacional, 18,* 69-110. Retirado de: http://www.fcc.org.br/pesquisa/publicacoes/eae/arquivos/1043/1043.pdf

Wittrock, M. C. (1992). An Empowering Conception of Educational Psychology. *Educational Psychologist, 27* (2), 129-141.

Zeferino, A. M. & Passeri, S. M. (2007). Avaliação da aprendizagem do estudante. *Cadernos ABEM, 3*, 39-43

ANEXO 1

Faculdade de Psicologia
UNIVERSIDADE DE LISBOA

MODELO DE UM ENUNCIADO DE EXAME (ESTRUTURA DE 2011/12) COM EXEMPLOS DE UM ITEM DE ESCOLHA MÚLTIPLA E DE UM ITEM DE DESENVOLVIMENTO.

PSICOLOGIA DIFERENCIAL

MIP. 1ºCiclo. 3º Ano. 2º Semestre

20 /20

Exame de ….ÉPOCA

Data , Hora , Salas

Nome:_____ **Nº**_____

1ª Parte: Questões de Escolha Múltipla

Por favor, em cada uma das seguintes 10 questões de escolha múltipla, assinale a alternativa de resposta que considera correta (apenas uma é completamente correta) e, de seguida, justifique a sua opção utilizando o espaço que se segue à questão. **Atenção**: deve justificar indicando por que razão considera a opção que escolheu correta e não se limitar a comentar em termos genéricos o tema da questão ou a justificar porque são as outras duas opções incorretas. **O limite de linhas disponíveis deve ser respeitado**.

1	As conceções sistémicas da inteligência, como a Teoria Triárquica de Sternberg, vêm opor-se à tradicional noção de inteligência geral ou de *fator g* devido…	
R: opção	**A**	…ao carácter interno, abstrato e limitado da gama de funcionamento intelectual que é contemplada por essa noção tradicional.
	B	…à incapacidade de os autores que defendem a noção clássica demonstrarem, sem margem para dúvidas, que o *fator g* emerge em todas as análises fatoriais no domínio cognitivo.
	C	…à diferenciação interindividual ser muito mais ampla em variáveis complexas do que em variáveis moleculares, como as tradicionais aptidões.
Justificação da opção:		

2	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

3	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

4	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção (Questão 4)**:**		

5	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

6	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

7	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

8	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

9	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção (Questão 9)**:**		

10	Enunciado (afirmação incompleta)…	
R: opção	**A**	…frase que completa a afirmação.
	B	…frase que completa a afirmação.
	C	…frase que completa a afirmação.
Justificação da opção:		

2ª Parte: Questão de Desenvolvimento

Desenvolva o tema que se segue, **não ultrapassando o limite de espaço** disponibilizado para resposta:

TEMA PARA DESENVOLVIMENTO:

Situe e caracterize a abordagem dimensional das diferenças individuais no quadro do desenvolvimento da Psicologia Diferencial e refira os seus contributos e limitações. Ilustre a sua resposta no domínio conativo referindo-se ao contraste entre as variáveis dimensionais e outros tipos de variáveis diferenciais.

Modelo de FCE – Ficha de Classificação de Exame

CLASSIFICAÇÃO – ALUNO Nº_____

QUESTÃO	RC	Cotação	R	Observações	Class.	Class. Acumul.
1		1				
2		1				
3		1				
4		1				
5		1				
6		1				
7		1				
8		1				
9		1				
10		1				
D		10				
				(TOTAL x .70 =)	**TOTAL**	

Avaliação da Resposta de Desenvolvimento:

ANEXO 2

Quadro 13 – Número e percentagem de examinados que selecionaram cada distrator no Grupo 1 (piores alunos) e Grupo 2 (melhores alunos), nos exames de Psi. Diferencial (três épocas), no ano letivo 2010/11.

Exame		Itens	Resposta Correta	Distratores					
				A		B		C	
				Grupo1	Grupo2	Grupo1	Grupo2	Grupo1	Grupo2
Psi.Diferencial (2010/11) -1ªépoca	Téorica (n=53, Q_1=31, Q_4=22)	1	C	11 (35,5%)	7 (31,8%)	7 (22,6%)	1 (4,5%)		
		2	B	7 (22,6%)	1 (4,5%)			12 (38,7%)	4 (18,2%)
		3	C	3 (9,7%)	3 (13,6%)	9 (29%)	2 (9,1%)		
		4	A			5 (16,1%)	1 (4,5%)	12 (38,7%)	0
		5	A			16 (51,6%)	8 (36,4%)	12 (38,7%)	0
	Prática (n= 60 , Q_1=37, Q_4= 23)	6	C	12 (32,4%)	5 (21,7%)	21 (56,8%9	2 (8,7%)		
		7	B	7 (18,9%)	2 (8,7%)			16 (43,2%)	9 (39,1%)
		8	B	14 (37,8%)	2 (8,7%)			8 (21,6%)	0
		9	C	9 (24,3%)	3 (13%)	12 (32,4%)	6 (26,1%)		
		10	A			14 (37,8%)	7 (30,4%)	9 (24,3%)	4 (17,4%)
Psi.Diferencial (2010/11) – 2ªépoca	Teórcia (n=53, Q_1=29, Q_4= 24)	1	B	11 (37,9%)	7 (29,2%)			12 (41,4%)	7 (29,2%)
		2	C	3 (10,3%)	0	14 (48,3%)	0		
		3	A			10 (34,5%)	2 (8,3%)	9 (31%)	0
		4	B	18 (62,1%)	6 (25%)			9 (31%)	9 (37,5%)
		5	A			13 (44,8%)	0	6 (20,7%)	0
	Prática (n= 55, Q_1=31, Q_4=24)	6	C	16 (51,6%)	1 (4,2%)	4 (12,9%)	1 (4,2%)		
		7	B	14 (45,2%)	0			13 (41,9%)	1 (4,2%)
		8	A			1 (3,2%)	0	14 (45,2%)	0
		9	C	22 (71%)	5 (20,8%)	6 (19,4%)	3 (12,5%)		
		10	B	7 (22,6%)	2 (8,3%)			16 (51,6%)	0
Psi.Diferencial (2010/11) - Ép.Esp.	Teórica (n=27 , Q_1=15,Q4= 12)	1	A			4 (26,7%)	1 (8,3%)	6 (40%)	4 (33,3%)
		2	C	6 (40%)	0	3 (20%)	0		
		3	B	3 (20%)	0			7 (46,7%)	1 (8,3%)
		4	C	4 (26,7%)	1 (8,3%)	3 (20%)	1 (8,3%)		
		5	A			7 (46,7%)	1 (8,3%)	4 (26,7%)	0
	Prática (n27= , Q_1=16,Q_4= 11)	6	A			7 (43,8%)	1 (9,1%)	2 (12,5%)	0
		7	B	7 (43,8%)	3 (27,3%)			8 (50%)	0
		8	B	6 (37,5%)	2 (18,2%)			6 (37,5%)	1 (9,1%)
		9	A			4 (25%)	4 (36,4%)	7 (43,8%)	3 (27,3%)
		10	C	6 (37,5%)	3 (27,3%)	5 (31,3%)	0		

Nota: Grupo 1 representa o Quartil 1(Q_1) e o Grupo 2, o Quartil 4 (Q_4), das distribuições das classificações no conjunto dos itens de escolha múltipla (EM) (critério proposto por DiBattisa e Kursawa, 2011).

Quadro 14 – Número e percentagem de examinados que selecionaram cada distrator no Grupo 1 (piores alunos) e Grupo 2 (melhores alunos), nos exames de Psi. Diferencial (três épocas), no ano letivo 2011/12.

Exame	Itens	Resposta Correta	Distratores					
			A		B		C	
			Grupo1	Grupo2	Grupo1	Grupo2	Grupo1	Grupo2
Psi.Diferencial (2011/12) -1ªépoca (n=88, Q_1= 48, Q_4= 40)	1	C	17 (35,4%)	7 (17,5%)	12 (25%)	0		
	2	B	15 (31,3%)	6 (15%)			15 (31,3%)	5 (12,5%)
	3	B	22 (45,8%)	2 (5%)			10 (20,8%)	3 (7,5%)
	4	C	11 (22,9%)	0	11 (22,9%)	1 (2,5%)		
	5	C	11 (22,9%)	0	11 (22,9%)	1 (2,5%)		
	6	B	19 (39,6%)	4 (10%)			13 (27,1%)	1 (2,5%)
	7	C	14 (29,2%)	1 (2,5%)	12 (25%)	0		
	8	A			34 (70,8%)	4 (10%)	4 (8,3%)	0
	9	A			24 (50%)	14 (35%)	20 (41,7%)	12 (30%)
	10	C	18 (37,5%)	6 (15%)	14 (29,2%)	2 (5%)		
Psi.Diferencial (2011/12) – 2ªépoca (n=57, Q_1= 32, Q_4= 25)	1	C	11 (34,4%)	10 (40%)	9 (28,1%)	4 (16%)		
	2	C	16 (50%)	3 (12%)	3 (9,4%)	1 (4%)		
	3	A			11 (34,4%)	5 (20%)	14 (43,8%)	6 (24%)
	4	B	7 (21,9%)	1 (4%)			11 (34,4%)	1 (4%)
	5	B	3 (9,4%)	0			16 (50%)	2 (8%)
	6	A			13 (40,6%)	1 (4%)	8 (25%)	0
	7	C	0	0	28 (87,5%)	11 (44%)		
	8	B	13 (40,6%)	2 (8%)			9 (28,1%)	3 (12%)
	9	C	7 (21,9%)	8 (32%)	15 (46,9%)	7 (28%)		
	10	A			5 (15,6%)	1 (4%)	12 (37,5%)	1 (4%)
Psi.Dierencialf. (2011/12) - Ép.Esp. (n= 21, Q_1= 11, Q_4= 10)	1	B	5 (45,5%)	1 (10%)			1 (9,1%)	1 (10%)
	2	C	4 (36,4%)	0	2 (18,2%)	0		
	3	B	2 (18,2%)	1 (10%)			0	0
	4	B	3 (27,3%)	1 (10%)			3 (27,3%)	1 (10%)
	5	A			5 (45,5%)	0	6 (54,5%)	0
	6	C	2 (18,2%)	0	8 (72,7%)	3 (30%)		
	7	B	1 (9,1%)	3 (30%)			3 (27,5%)	0
	8	C	4 (36,4%)	1 (10%)	4 (36,4%)	0		
	9	A			4 (36,4%)	2 (20%)	5 (45,5%)	2 (20%)
	10	A			4 (36,4%)	1 (10%)	6 (54,5%)	0

Nota: Grupo 1 representa o Quartil 1(Q_1) e o Grupo 2, o Quartil 4 (Q_4), das distribuições das classificações no conjunto dos itens de escolha múltipla (EM) (critério proposto por DiBattisa e Kursawa, 2011).

Quadro 15 – Número e percentagem de examinados que selecionaram cada distrator no Grupo 1 (piores alunos) e Grupo 2 (melhores alunos) nos exames de Psicometria (três épocas), no ano letivo 2011/12.

Exame	Itens	Resposta Correta	Distratores					
			A		B		C	
			Grupo1	Grupo2	Grupo1	Grupo2	Grupo1	Grupo2
Psicometria (2011/12) -1ªépoca (n= 74,Q_1= 38, Q_4= 36)	1	C	6 (15,8%)	0	8 (21,1%)	0		
	2	C	14 (36,8%)	1 (2,8%)	3 (7,9%)	0		
	3	B	6 (15,8%)	0			4 (10,5%)	3 (8,3%)
	4	B	14 (36,8%)	1 (2,8%)			9 (23,7%)	6 (16,7%)
	5	C	11 (28,9%)	2 (5,6%)	19 (50%)	22 (61%)		
	6	C	22 (57,9%)	3 (8,3%)	10 (26,3%)	2 (5,6%)		
	7	B	12 (31,6%)	0			13 (34,2%)	4 (11,1%)
	8	C	13 (34,2%)	3 (8,3%)	10 (26,3%)	0		
	9	A			12 (31,6%)	3 (8,3%)	14 (36,8%)	4 (11,1%)
	10	C	6 (15,8%)	2 (5,6%)	10 (26,3%)	0		
Psicometria (2011/12) - 2ªépoca (n= 58,Q_1= 30, Q_4= 28)	1	C	10 (33,3%)	6 (21,4%)	13 (43,3%)	7 (25%)		
	2	B	27 (90%)	18 (64,3%)			1 (3,3%)	0
	3	B	8 (26,7%)	2 (7,1%)			15 (50%)	3 (10,7%)
	4	C	8 (26,7%)	1 (3,6%)	4 (13,3%)	1 (3,6%)		
	5	A			1 (3,3%)	1 (3,6%)	11 (36,7%)	2 (7,1%)
	6	C	6 (20%)	1 (3,6%)	15 (50%)	7 (25%)		
	7	A			7 (23,3%)	0	6 (20%)	1 (3,6%)
	8	A			9 (30%)	2 (7,1%)	12 (40%)	3 (10,7%)
	9	C	6 (20%)	5 (17,9%)	14 (46,7%)	0		
	10	B	3 (10%)	0			16 (53,3%)	8 (28,6%)
Psicometria (2011/12) - Ép.Esp. (n= 12, Q_1= 13, Q_4= 12)	1	A			4 (30,5%)	1 (8,3%)	2 (15,4%)	0
	2	C	1 (7,7%)	0	0	0		
	3	B	8 (61,5%)	1 (8,3%)			3 (23,1%)	0
	4	A			8 (61,5%)	3 (25%)	3 (23,1%)	0
	5	C	7 (53,8%)	8 (66,7%)	4 (30,8%)	0		
	6	A			7 (53,8%)	1 (8,3%)	0	0
	7	C	1 (7,7%)	0	8 (61,5%)	0		
	8	C	5 (38,5%)	0	1 (7,7%)	0		
	9	B	4 (30,8%)	0			4 (30,8%)	0
	10	A			2 (15,4%)	0	1 (7,7%)	0

Nota: Grupo 1 representa o Quartil 1(Q_1) e o Grupo 2, o Quartil 4 (Q_4), das distribuições das classificações no conjunto dos itens de escolha múltipla (EM) (critério proposto por DiBattisa e Kursawa, 2011).

Quadro 16 - Médias, desvios-padrão, mínimos e máximos, nºde exames, dos examinados repetentes que realizaram os dois formatos de exames de Psi. Diferencial, em dois anos letivos, 2010/11 e 2011/12

Examinados (n=32)	Nº de exames	Média Total (0 – 20)	Desvio-padrão	Min.	Máx.	Média do 1ºformato (PD, 10-11) (0 – 20)	Média do 2ºformato (PD, 11-12) (0 – 20)
1	4	4,94	1,56	3,25	6,50	3,25	5,50
2	3	5,92	4,26	1,50	10,00	1,50	8,13
3	3	8,33	1,04	7,50	9,50	7,50	8,75
4	3	6,00	,63	6,00	7,25	6,63	6,75
5	5	5,85	1,59	4,25	7,75	5,50	7,13
6	5	7,40	2,10	4,25	10,00	6,67	9,00
7	4	8,06	1,14	6,75	9,50	7,25	8,88
8	3	7,50	3,03	5,75	11,00	5,75	8,38
9	3	2,67	1,13	1,50	3,75	1,50	3,25
10	5	6,90	2,26	4,75	10,50	6,33	9,50
11	5	5,15	,76	4,25	6,00	4,38	5,67
12	5	5,80	2,87	3,50	9,50	4,38	7,33
13	4	6,13	3,26	3,00	10,75	4,75	8,00
14	4	6,19	1,80	3,50	7,25	5,25	7,13
15	5	6,15	3,01	1,25	9,25	6,08	6,25
16	3	7,83	2,38	7,50	11,75	7,50	9,75
17	4	4,94	1,95	2,50	7,25	4,75	5,00
18	5	4,85	1,55	3,00	6,75	3,38	5,83
19	4	5,63	1,79	3,75	8,00	5,88	5,38
20	5	6,35	2,94	1,50	9,50	7,00	5,92
21	4	5,69	1,38	3,75	7,00	4,88	6,50
22	3	5,25	2,84	4,00	9,50	5,50	6,75
23	5	6,00	2,84	2,00	9,50	3,50	8,00
24	4	8,19	2,41	6,50	11,75	7,00	9,38
25	4	3,25	1,70	1,25	4,75	2,00	4,75
26	4	2,06	,94	1,00	3,25	1,75	2,63
27	3	6,25	5,07	,75	10,75	,75	9,00
28	4	7,13	1,76	4,50	8,25	4,50	8,00
29	3	7,33	5,35	4,00	13,50	4,50	8,75
30	2	6,00	3,54	3,75	8,75	3,75	8,75
31	3	7,42	2,43	4,75	9,50	4,75	8,75
32	3	8,42	3,22	5,75	12,00	5,75	9,75

Quadro 17 – Teste de Wilcoxon, para amostras emparelhadas

		N	Média	Soma dos Ranks
Média de PD(11/12), 1ªép., 2ªép. e ép.esp – Media de PD (10/11), 1ªép., 2ªép. e ép. Esp	Ranks Negativos	2[a]	5,00	10,00
	Ranks Positivos	30[b]	17,27	518,00
	Empates	0[c]		
	Total	32		

a. Média PD(11/12), 1ªép., 2ªép., ép.esp. < Média PD(10/11) 1ªép., 2ªép., ép.esp

b. Média PD(11/12), 1ªép., 2ªép., ép.esp. > Média PD(10/11) 1ªép., 2ªép., ép.esp

c. Média PD(11/12), 1ªép., 2ªép., ép.esp. = Média PD(10/11) 1ªép., 2ªép., ép.esp

Nota: Estatística de Teste, Z = -4.75 ($p<.00$)

Quadro 18 – Média total, desvio-padrão, mínimos e máximos, nº de exames, dos examinados repetentes que realizaram o exame de Psicometria em 2011/12

Examinados	Nº de exames	Média Total	Desvio-padrão	Min.	Máx.
1	3	7,83	2,13	6,25	10,25
2	2	6,75	1,77	5,50	8,00
3	2	2,75	1,06	2,00	3,50
4	2	6,88	,88	6,25	7,50
5	2	8,38	2,30	6,75	10,00
6	2	6,25	4,60	3,00	9,50
7	3	8,83	4,07	6,00	13,50
8	3	4,58	1,38	3,25	6,00
9	3	7,00	,75	6,25	7,75
10	3	8,42	3,55	5,25	12,25
11	2	9,38	1,94	8,00	10,75
12	2	5,75	,71	5,25	6,25
13	3	8,75	4,02	5,50	13,25
14	2	6,38	,18	6,25	6,50
15	2	7,13	7,95	1,50	12,75
16	3	6,25	1,75	4,50	8,00
17	2	9,63	3,36	7,25	12,00
18	3	9,42	2,93	7,25	12,75
19	2	6,00	1,41	5,00	7,00
20	2	7,88	4,07	5,00	10,75
21	2	5,50	3,54	3,00	8,00
22	2	9,75	1,77	8,50	11,00
23	2	10,13	1,94	8,75	11,50
24	2	10,63	2,65	8,75	12,50
25	3	5,67	1,28	4,25	6,75
26	2	4,13	1,24	3,25	5,00
27	3	8,42	2,40	6,25	11,00
28	2	8,75	3,54	6,25	11,25
29	3	5,42	1,42	4,25	7,00
30	2	9,75	4,60	6,50	13,00
31	2	7,33	1,28	6,25	8,75

32	2	8,63	3,01	6,50	10,75
33	3	8,83	4,07	6,00	13,50
34	3	6,58	3,39	3,00	9,75
35	3	3,75	1,98	2,25	6,00
36	3	7,58	2,02	5,75	9,75
37	2	7,50	4,60	4,25	10,75
38	2	5,00	1,77	3,75	6,25
39	2	5,50	,00	5,50	5,50
40	3	8,58	4,25	4,25	12,75
41	2	12,13	4,77	8,75	15,50
42	3	6,67	1,15	6,00	8,00
43	2	4,88	,88	4,25	5,50
44	3	9,33	5,62	4,50	15,50
45	3	5,25	2,17	4,00	7,75
46	3	8,83	6,15	3,00	15,25
47	2	9,00	2,83	7,00	11,00
48	2	9,88	3,71	7,25	12,50
49	2	4,00	2,12	2,50	5,50
50	2	4,50	2,47	2,75	6,25
51	3	7,50	3,47	5,25	11,50
52	2	9,25	2,47	7,50	11,00

Quadro 19 – Média, desvio-padrão, variância e correlação entre o número de exames efetuados e a média das classificações dos estudantes que realizara mais do que um exame de Psi. Diferencial no mesmo ano letivo e em ambos os anos letivos

	Psicologia Diferencial (2010/11) (N=82)		Psicologia Diferencial (2011/12) (N=78)		Psicologia Diferencial (2010/11 + 2011/12) (N=71)	
	Nº exames efetuados	Média das classificações	Nº exames efetuados	Média das classificações	Nº exames efetuados	Média das classificações
Amplitude	2 – 3	1.75 – 14.25	2 – 3	1.83 – 12.75	2 – 5	2.06 – 10.63
Media	2.22	7.82	2.24	7.64	2.80	6.73
Desvio-padrão	.42	2.48	.43	2.10	1.08	1.7
Variância	.17	6.15	.19	4.43	1.61	2.46
Correlação de Spearman entre as duas variáveis	-.13		-.40		-.37	